Sveučilište Josipa Jurja Strossmayera u Osijeku

Odjel za kemiju

PRIRUČNIK ZA SEMINAR IZ KEMOMETRIČKIH METODA

Vlatka Gvozdić
Elvira Kovač-Andrić

2026.

Izdavač
SVEUČILIŠTE JOSIPA JURJA STROSSMAYERA U OSIJEKU
ODJEL ZA KEMIJU

Za izdavača:
izv. prof. dr. sc. Elvira Kovač-Andrić

Autorice:
izv. prof. dr. sc. Vlatka Gvozdić
izv. prof. dr. sc. Elvira Kovač-Andrić

Recenzenti:
izv. prof. dr. sc. Lidija Kalinić
izv. prof. dr. sc. Domagoj Vidosavljević

Lektor:
izv. prof. dr. sc. Borko Baraban

veljača, 2026. godine

ISBN 978-953-8154-37-9
CIP zapis je dostupan u nacionalnom skupnom katalogu knjižničnog sustava Bukinet pod brojem 991005957104909366

Sveučilište Josipa Jurja Strossmayera u Osijeku
Suglasnost za izdavanje ovog sveučilišnog priručnika donio je Senat
Sveučilišta Josipa Jurja Strossmayera u Osijeku
na 6. sjednici održanoj 31. ožujka 2026. godine pod brojem 6/26.

Sadržaj

PREDGOVOR

Kemometričke metode danas predstavljaju nezaobilazno sredstvo u obradi, analizi i interpretaciji složenih kemijskih podataka. Razvoj suvremenih analitičkih tehnika rezultirao je velikim količinama podataka čija obrada zahtijeva primjenu statističkih i matematičkih metoda prilagođenih kemijskim sustavima. U cilju približavanja tih metoda studentima diplomskog studija kemije, nastao je ovaj priručnik.

U priručniku su opisane temeljne kemometričke metode, uključujući regresijsku analizu, analize vremenskih serija, diskriminacijsku analizu, analizu glavnih komponenti i klastersku analizu. Naglasak je stavljen na praktičnu primjenu i razumijevanje metoda s pomoću konkretnih eksperimentalnih primjera.

Priručnik je koncipiran problemski pri čemu su poglavlja praćena kratkim teorijskim uvodom, riješenim zadatcima i pitanjima za provjeru znanja. Cilj je priručnika pružiti studentima kemije temeljno razumijevanje odabranih multivarijantnih metoda obrade podataka i njihovih teorijskih pretpostavki te područja primjene. Poseban naglasak stavljen je na interpretaciju rezultata u kontekstu kemijskih eksperimenata čime se postiže razvoj kritičkog mišljenja pri odabiru najpogodnije metode obrade podataka.

Ovaj priručnik, iako namijenjen studentima kemije, može biti koristan i studentima drugih znanstvenih područja (biologija, medicina, farmacija) koji se u svojem radu susreću s multivarijantnim metodama analize podataka. Nadamo se da će doprinijeti lakšem razumijevanju kemometričkih metoda i potaknuti njihovu primjenu u akademskom i istraživačkom radu.

Autorice

1. REGRESIJSKA ANALIZA

Regresijska je analiza iznimno važna u kemijskim istraživanjima jer omogućuje kvantitativno opisivanje odnosa među eksperimentalno dobivenim varijablama. U kemiji se regresijski modeli najčešće primjenjuju za ispitivanje ovisnosti fizikalno-kemijskih svojstava o sastavu sustava, reakcijskim uvjetima ili strukturnim značajkama molekula kao i u analizi i interpretaciji eksperimentalnih podataka.

Primjena regresijske analize osobito je važna u analitičkoj kemiji, fizikalnoj kemiji, u području kemometrije i kemijskom inženjerstvu. Primjeri obuhvaćaju određivanje kalibracijskih pravaca u instrumentalnim metodama analize, modeliranje kinetike kemijskih reakcija, procjenu utjecaja temperature i koncentracije na brzinu reakcije kao i uspostavljanje odnosa strukture i svojstava tvari poznatih kao QSAR i QSPR modeli (engl. *Quantitative Structure-Activity Relationship*, QSAR; engl. *Quantitative Structure-Property Relationship*, QSPR).

Najčešće upotrebljavani regresijski model u kemijskim primjenama jest linearna regresija, osobito u postupcima kalibracije, gdje se pretpostavlja linearna ovisnost signala instrumenta o koncentraciji analizirane tvari. Kada sustav obuhvaća više nezavisnih varijabli, primjenjuje se višestruka linearna regresija, dok se u slučajevima nelinearnog ponašanja kemijskih sustava upotrebljavaju nelinearni regresijski modeli.

Ključan dio regresijske analize u kemiji čini procjena parametara modela i ocjena njihove pouzdanosti, najčešće metodom najmanjih kvadrata. Jednako je važno provjeriti pretpostavke modela, uključujući linearnost, homogenost varijance i neovisnost pogrešaka, kako bi se osigurale točnost i ponovljivost rezultata. Nepravilna primjena regresijske analize može dovesti do pogrešne interpretacije podataka i netočnih zaključaka.

U kemiji se uspješnost regresijskog modela ne procjenjuje na temelju jednog pokazatelja, već kombinacijom statističkih kriterija, analize rezidualnih vrijednosti i kemijskog smisla modela, s posebnim naglaskom na sposobnost predviđanja i eksperimentalnu ponovljivost.

1.1. Linearna regresija

Regresijski je model jednadžba s konačnim brojem parametara i varijabli, a linearna regresijska jednadžba predstavljena je izrazom (1):

$$y = a + b \cdot x \qquad (1).$$

Veličina *a* jest konstantni član (odsječak na ordinati), a *b* je nagib pravca. Konstantni član *a* predstavlja vrijednost regresijske funkcije ako je vrijednost neovisne varijable (*x*) jednaka nuli. Regresijski koeficijent *b* predstavlja iznos linearne promjene regresijske vrijednosti pri jediničnoj promjeni vrijednosti neovisne varijable *x.*

1.1.1. Pokazatelji uspješnosti regresijskog modela

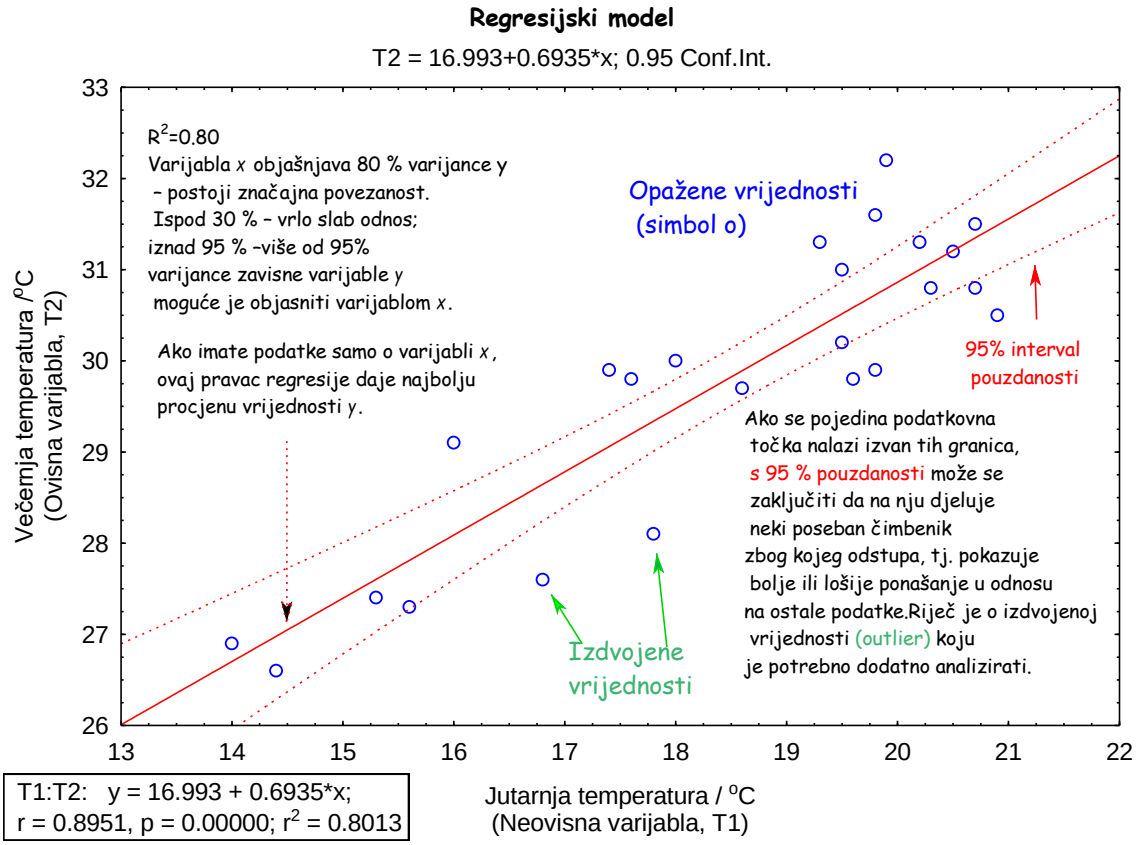
Pokazatelji su uspješnosti regresijskog modela: koeficijent determinacije (R^2, vrijednosti od 0 do 1), prilagođeni koeficijent determinacije ($R_{adj,}$ vrijednosti od 0 do 1), standardna pogreška procjene (engl. *Standard Error of Estimate*, SE) i korijen srednje kvadratne pogreške (engl. *Root Mean Squared Error*, RMSE), rezidualne vrijednosti, statistička značajnost parametara (t-test, p-vrijednost), F-test (ukupna značajnost modela), granice pouzdanosti (*Confidence limits*) i interval pouzdanosti (*Confidence interval).*

Visoka vrijednost R^2 ne pokazuje nužno ispravnost modela ni njegovu sposobnost predviđanja. Prilagođeni koeficijent determinacije (R_{adj}) upotrebljava se prvenstveno kod višestruke regresije, a sprječava umjetno povećavanje R^2 uslijed dodavanja novih varijabli koje nisu relevantne.

Standardna pogreška procjene mjeri prosječno odstupanje eksperimentalnih vrijednosti od vrijednosti predviđenih modelom. Niže vrijednosti upućuju na precizniji model, a u kemiji se često izražava istim jedinicama kao mjerena veličina (npr. množinska koncentracija, mol/L). Standardna pogreška procjene slična je korijenu srednje kvadratne pogreške, ali koristi stupnjeve slobode ($n - k - 1$) umjesto n.

Rezidualne vrijednosti predstavljaju razliku između eksperimentalnih i modelom predviđenih vrijednosti. Nasumična raspodjela rezidualnih vrijednosti i približno normalna raspodjela upućuju

na dobar model, dok sustavni trendovi mogu upućivati na nelinearnost, heteroskedastičnost ili pogreške u mjerenju. Statistička značajnost parametara (t-test, p-vrijednost) procjenjuju značajnost regresijskih koeficijenata. Obično se smatraju statistički značajnim ako je $p \leq 0{,}05$. F-test ispituje statističku značajnost cijelog regresijskog modela. U regresijskoj su analizi Q^2 (engl. *predictive R-squared*, Q^2) i PRESS (engl. *Predicted Residual Sum of Squares*, PRESS) mjere koje upućuju na sposobnost predviđanja nekog modela, odnosno upućuju na sposobnost modela da točno predvidi vrijednosti unutar novih, neovisnih skupova podataka. Važni su u kemometričkim metodama i QSAR/QSPR modelima. Niže vrijednosti PRESS i više vrijednosti Q^2 znače bolju moć predviđanja modela. U regresijskoj analizi interval pouzdanosti pokazuje raspon u kojem se s određenom sigurnošću nalazi stvarna vrijednost regresijskog koeficijenta, a granice pouzdanosti predstavljaju donju i gornju granicu tog raspona (slika 1.).



Slika 1. Rezultat regresijske analize. Dijagram prikazuje ovisnost večernje temperature zraka (T_2) o jutarnjoj temperaturi (T_1) (program *Statistica*).

1.2. Višestruka regresijska analiza

Višestruka regresijska analiza statistička je metoda koja omogućuje istovremeno modeliranje utjecaja više različitih čimbenika na jednu pojavu. Za razliku od jednostavne regresije koja promatra odnos između dviju varijabli, višestruka regresija omogućuje uključivanje više nezavisnih varijabli što daje realniji opis stvarnih sustava.

U kemiji se ta metoda često primjenjuje jer kemijski procesi rijetko ovise o samo jednom čimbeniku – primjerice, brzina kemijske reakcije može ovisiti o temperaturi, koncentraciji reaktanata, tlaku i prisutnosti katalizatora. Višestruka regresijska analiza omogućuje procjenu utjecaja svih tih čimbenika i ujedno predviđa ponašanje sustava u različitim uvjetima.

Primjenom višestruke regresije kemičari mogu analizirati eksperimentalne podatke, optimizirati uvjete reakcija, poboljšati točnost predviđanja i bolje razumjeti složene kemijske odnose. Jednadžba je višestruke regresije (2):

$$y = a + b_1x_1 + b_2x_2 + b_3x_3.... + b_nx_n \qquad (2)$$

gdje su x_1, x_2, x_3 nezavisne varijable, a b_1, b_2 *i* b_3 regresijski koeficijenti koji opisuju utjecaj svake varijable.

Na primjer, promatra li se koncentraciju troposferskog ozona u ovisnosti o temperaturi zraka (*T*), atmosferskom tlaku (*p*) i brzini vjetra (*BV*), uvrštavanjem tih varijabli u model višestruke regresije dobiva se izraz (3):

$$c\,(O_3) = a + b_1T + b_2p + b_3BV \qquad (3)$$

u kojem koeficijenti b_1, b_2 *i* b_3 pokazuju koliko se koncentracija troposferskog ozona (O_3) mijenja pri promjeni svakog pojedinog čimbenika. Takva interpretacija omogućuje izolaciju pojedinačnog doprinosa svakog okolišnog čimbenika na ukupnu koncentraciju ozona.

Kao i kod jednostavne linearne regresije, uspješnost višestruke regresijske analize procjenjuje se s pomoću nekoliko statističkih pokazatelja. Najvažniji je koeficijent determinacije (R^2) koji pokazuje koliki dio varijabilnosti zavisne varijable model može objasniti. Budući da se vrijednost R^2 povećava s brojem nezavisnih varijabli, često se koristi i prilagođeni koeficijent determinacije (R^2_{adj}) koji uzima u obzir složenost modela. Dodatno, statistička značajnost modela ispituje se F-testom, dok se utjecaj pojedinih varijabli procjenjuje *t*-testom regresijskih koeficijenata. Analiza rezidualnih vrijednosti služi za provjeru odstupanja između eksperimentalnih i predviđenih vrijednosti kao i za procjenu pouzdanosti modela. Dobri modeli imaju nasumce raspoređene rezidualne vrijednosti, približno normalne raspodjele.

Primjer: Kalibracijski pravac

Kalibracijski pravac za određivanje koncentracije analizirane tvari proveden je mjerenjem signala standardnih otopina poznatih koncentracija (c_1, c_2, c_3, c_4 i c_5). Analitički signal pri nultoj koncentraciji trebao bi biti jednak nuli. Pregledom regresijskog modela potrebno je utvrditi da odsječak kalibracijskog pravca nije statistički značajan, a uvjet je da pripadajuća p-vrijednost bude veća od odabrane razine značajnosti ($p \geq 0.05$).

S obzirom na to, model se pojednostavljuje, a kalibracijski pravac opisuje jednadžbom (4):

$$y = b \cdot x \qquad (4)$$

gdje je y analitički signal, x koncentracija tvari, a b nagib pravca. Takav pristup omogućava jednostavnije i preciznije određivanje koncentracije nepoznatih uzoraka, uz zadržavanje statističke pouzdanosti modela.

1.3. **Zadatci**

1.3.1. Koristeći se podatcima iz tablice 1., izračunajte koncentraciju nepoznatog uzorka čija izmjerena vrijednost *A* = 1,612 pri valnoj duljini maksimuma a λmax =525 nm.

Tablica 1. Vrijednosti apsorbancije pri zadanim koncentracijama otopine $KMnO_4$

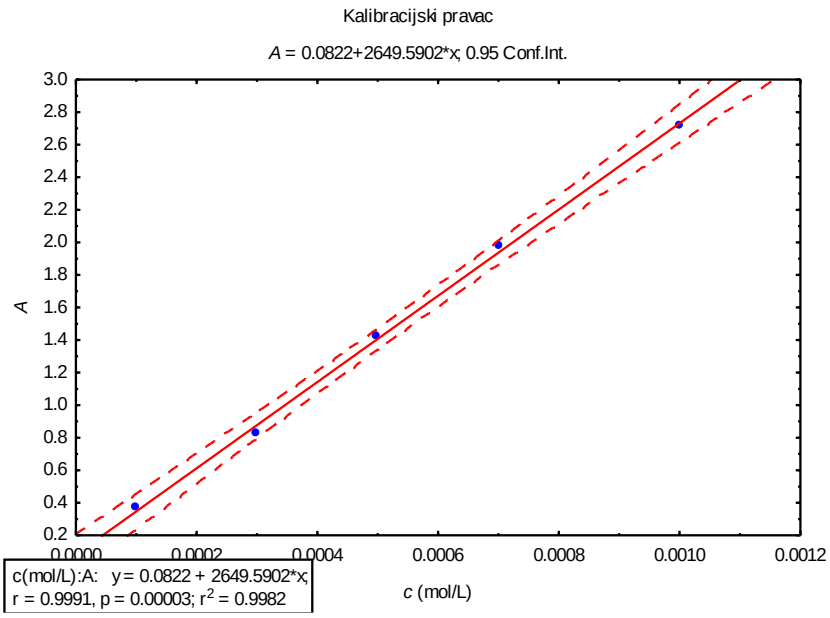
c (mol/L)	0,0001	0,0003	0,0005	0,0007	0,001
apsorbancija, *A*	0,37	0,82	1,42	1,90	2,71

Rezultati analize dobiveni obradom podataka u programskom paketu *Statistica* prikazani su u tablici 2., a grafička evaluacija modela prikazana je na slici 2. Kako bi se potvrdila preciznost dobivenih rezultata, na slici 3. prikazana je i usporedna evaluacija modela izrađena u programu *Microsoft Excel*.

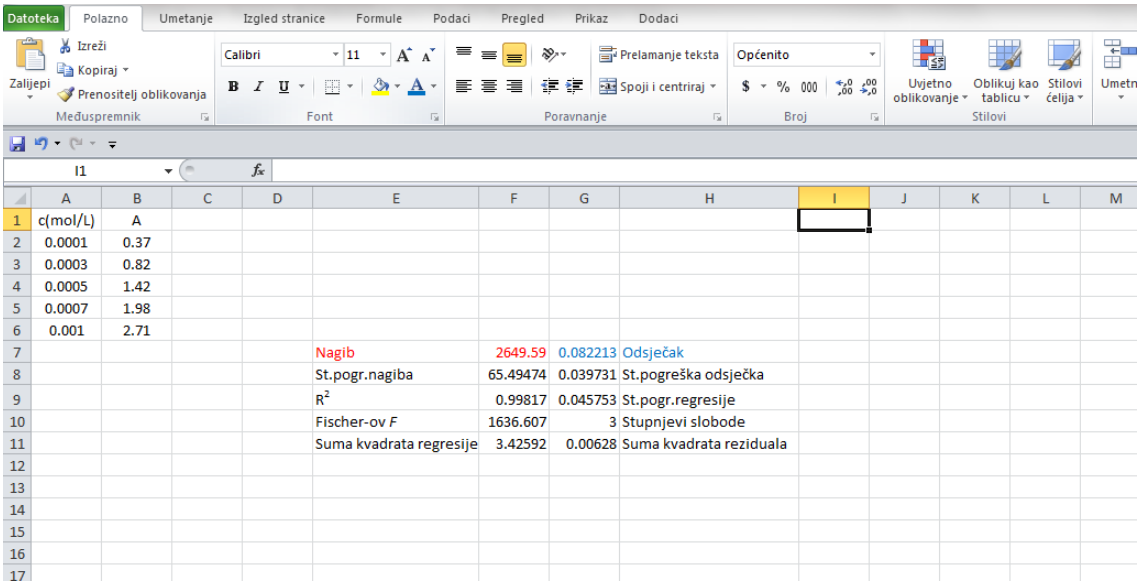
Tablica 2. Statistički parametri regresijske analize (program *Statistica*)

Regression Summary for Dependent Variable: *A* (Spreadsheet1) R= .99908473 R2= .99817029 Adjusted R2= .99756039 F(1,3)=1636.6 p						
	Beta	**Std.Err. - of Beta**	**B**	**Std.Err. - of B**	**t(3)**	**p-level**
Intercept			0.082	0.03973	2.06924	0.130337
***c*(mol/L)**	0.999085	0.024696	2649.590	65.49474	40.45500	0.000033

S obzirom na to da odsječak (engl. *intercept*) nije statistički značajan, pristupa se računanju po pojednostavljenom modelu (jednadžba 4).



Slika 2. Kalibracijski pravac s pripadajućim vrijednostima regresijskih koeficijenata i odgovarajućom jednadžbom (program *Statistica*)



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	c(mol/L)	A											
2	0.0001	0.37											
3	0.0003	0.82											
4	0.0005	1.42											
5	0.0007	1.98											
6	0.001	2.71											
7					Nagib	2649.59	0.082213	Odsječak					
8					St.pogr.nagiba	65.49474	0.039731	St.pogreška odsječka					
9					R^2	0.99817	0.045753	St.pogr.regresije					
10					Fischer-ov *F*	1636.607	3	Stupnjevi slobode					
11					Suma kvadrata regresije	3.42592	0.00628	Suma kvadrata reziduala					
12													
13													
14													
15													
16													
17													

Slika 3. Primjer sličnog izračuna u programu *Microsoft Excel*

S pomoću dobivenih rezultata prikazanih na slikama 2. i 3. izračunajte koncentraciju nepoznatog uzorka ako mu je izmjerena vrijednost apsorbancije pri λ_{max} = 525 nm iznosila *A* = 1,612. Zadatak riješite i grafički.

1.3.2. Clausius-Clapeyronova jednadžba povezuje tlak pare i temperaturu u faznim prijelazima ili pri promjeni stanja. Ako je poznata vrijednost tlaka pri više odabranih vrijednosti temperature, za izračun vrijednosti entalpije koristi se linearizirani oblik:

$$lnp = -\frac{\Delta H^0 vap}{R} \times \frac{1}{T} + konst. \qquad (5).$$

Na temelju eksperimentalnih podataka o ovisnosti tlaka pare o temperaturi, prikazanih u tablici 3., potrebno je odrediti standardnu entalpiju sublimacije joda (ΔH^0_{sub}) prema jednadžbi (6):

$$I_2(s) \rightarrow I_2(g) \qquad (6).$$

Izračunajte vrijednost ΔH^0_{sub} koristeći se nagibom pravca dobivenim regresijskom analizom navedenih mjerenja.

Tablica 3. Eksperimentalne vrijednosti temperature i tlaka pare tijekom procesa sublimacije joda.

$T_{subl.}$	271	281	291	300	310	320	330	341
$p(I_2)$/Pa	51,01	134,02	335,00	788,01	1756,00	3723,01	7543,03	14659,04

1.3.3. Na temelju eksperimentalnih podataka za reakciju $Na_2S_2O_3$ i HCl (tablica 4.) i dobivene vrijednosti nagiba pravca, izračunajte energiju aktivacije (*E*a) koristeći se linearnim oblikom Arrheniusove jednadžbe (7):

$$lnk = -\frac{E_a}{R}\left(\frac{1}{T}\right) + lnA \qquad (7).$$

Tablica 4. Eksperimentalne vrijednosti konstante brzine reakcije (*k*) pri različitim termodinamičkim temperaturama

t/℃	0	24	35	45	54	65
k/s^{-1}	$7{,}871 \times 10^{-7}$	$3{,}482 \times 10^{-5}$	$1{,}343 \times 10^{-4}$	$4{,}991 \times 10^{-4}$	$1{,}511 \times 10^{-3}$	$4{,}851 \times 10^{-3}$

1.3.4. Enzim katalaza katalizira razgradnju H_2O_2 u vodu i kisik. Pri koncentraciji enzima [E_0]= 2,41 nmol/L i pH = 7,8 dobiveni su rezultati prikazani u tablici 5. Odredite katalitičku aktivnost enzima (jednadžbe 9, 10), odnosno parametre: maksimalnu brzinu (V_{max}) i Michaelisovu konstantu (V_{max}), primjenom linearizacije po Lineweaveru i Burku (jednadžba 8):

$$\frac{1}{V_0} = \frac{K_m}{V_{max}}\left[\frac{1}{S}\right] + \frac{1}{V_{max}} \qquad (8)$$

$$k_{cat} = \frac{v_{max}}{[E_0]} \qquad (9)$$

$$katalitička\ aktivnost\ enzima = \frac{k_{cat}}{K_m} \qquad (10).$$

Tablica 5. Eksperimentalno određene početne brzine reakcije pri različitim koncentracijama

[H_2O_2] mmol/L	1,26	2,56	5,00	20,10
Brzina/mmol/Ls	0,0277	0,0502	0,0832	0,1661

1.3.5. Reakcija raspada H_2O_2 u prisutnosti male koncentracije Fe^{2+} katalitička je reakcija prvog reda. Izračunajte konstantu brzine reakcije (k_1) i $[A]_0$ na temelju eksperimentalnih rezultata prikazanih u tablici 6. Za precizno određivanje kinetičkih parametara koristi se linearizirani oblik integriranog zakona brzine za reakcije prvog reda (11):

$$[\ln A]_t = -k_1 t + \ln[A]_0 \qquad (11).$$

Tablica 6. Promjena koncentracije H_2O_2 u ovisnosti o vremenu tijekom katalitičkog raspada

t/s	2	4	6	8	10	12	14	16	18	20
[H_2O_2]mol/L	6,222	4,831	3,780	3,202	2,611	2,171	1,861	1,501	1,282	1,013

1.3.6. Proces racemizacije L-alanina s pomoću enzima alanin racemaze slijedi kinetiku drugog reda u odnosu na koncentraciju reaktanta, a podatci racemizacije prikazani su u tablici 7. Izračunajte

konstantu brzine reakcije na temelju danih podataka. Za određivanje kinetičkih parametara takvih sustava primjenjuje se integrirani oblik zakona brzine za reakcije drugog reda (12):

$$\frac{1}{[A]_t} = \boldsymbol{k_2 t} + \frac{1}{[A]_0} \qquad (12).$$

Tablica 7. Vrijednosti koncentracije L-alanina u ovisnosti o vremenu

t/s	0	601	1201	1802	2400
[A] mol/L	0,4110	0,3601	0,3121	0,2782	0,2552

1.3.7. U tablici 8. prikazani su prosječni mjesečni podatci koncentracija stratosferskog ozona, atmosferskog tlaka i temperature zraka na području Brazila za 2021. godinu. Napravite model koji pokazuje ovisnost koncentracija O_3 o atmosferskom tlaku i temperaturi. Komentirajte prikazane rezultate koji su prikazani u tablici 9. i slikama 4. i 5.

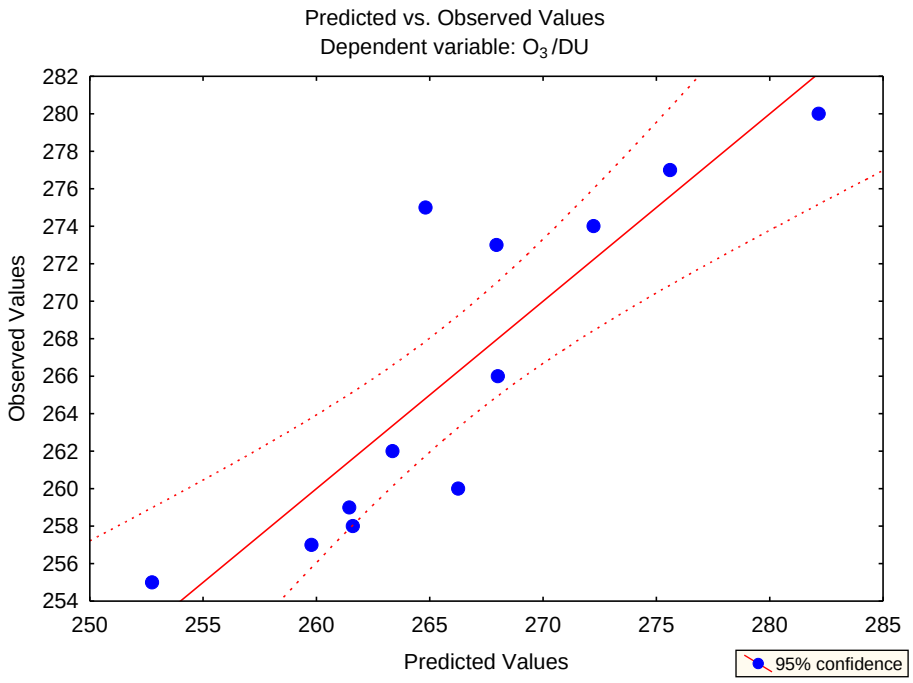
Tablica 8. Prosječni mjesečni podatci koncentracije stratosferskog ozona, atmosferskog tlaka i temperature zraka na području Brazila tijekom 2021. godine

Mjesec	1	2	3	4	5	6	7	8	9	10	11	12
O_3/DU	259	258	260	257	255	262	266	274	280	277	273	275
P/hPa	2,54	2,51	2,44	2,13	1,72	1,64	1,44	1,50	1,78	2,08	2,20	2,45
t/˚C	31,2	31,1	31,9	28,9	25,3	27,5	27,7	29	32,7	32,5	31,2	31,6

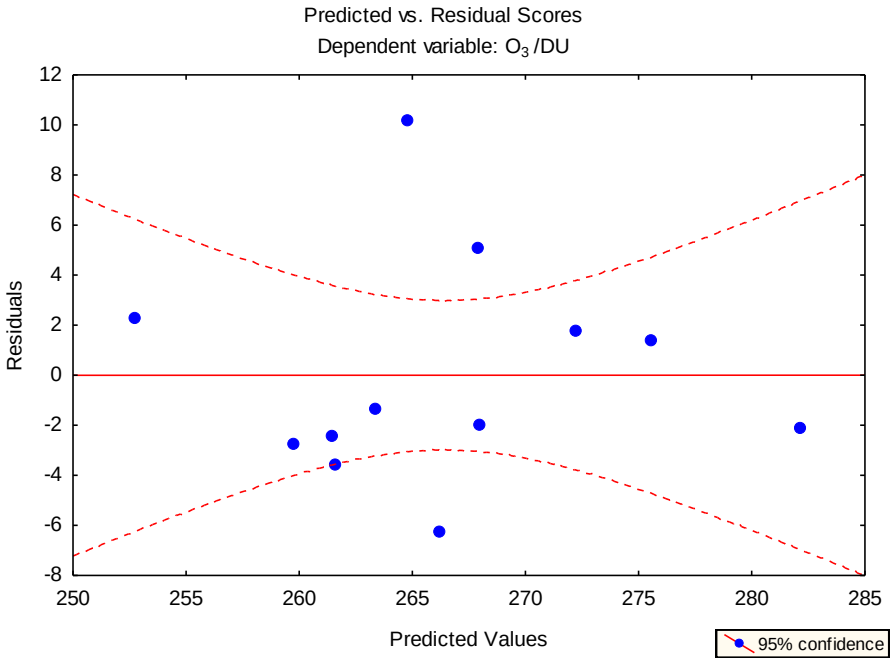
Rezultat provedene regresijske analize prikazan je u tablici 9. i slikama 4. i 5.

Tablica 9. Rezultat regresijske analize (program *Statistica*)

Regression Summary for Dependent Variable: O_3 DU (Brazil 2 mjesečni prosjeci) R = .86696051 R^2= .75162053 Adjusted R^2= .69642509 F(2,9)=13.617 p						
	Beta	**Std.Err. - of Beta**	**B**	**Std.Err. - of B**	**t(9)**	**p-level**
Intercept			**181.6938**	19.92031	9.0421	0.000008
Pressure	-0.869032	0.210757	**-19.0955**	4.63104	-4.12338	0.002585
Temperature	1.076446	0.210757	**4.1281**	0.80824	5.10751	0.000639



Slika 4. Odnos opaženih i predviđenih vrijednosti stratosferskog ozona. (program *Statistica*)



Slika 5. Raspodjela rezidualnih vrijednosti (*program Statistica*)

Pitanja:

a) Što je ovisna, a što neovisna varijabla?
b) Kako se interpretira koeficijent regresije?
c) Što označava konstanta u regresijskoj jednadžbi?
d) Definirajte koeficijent determinacije.
e) Ako je vrijednost R^2 kalibracijske krivulje npr. 0,89, može li se (*smije li se*) iz takvog pravca odrediti koncentracija nepoznatog uzorka?
f) Kako se regresijska analiza može koristiti u praćenju kinetike reakcije?
g) Na temelju vrijednosti koeficijenta determinacije (R^2) te grafičkog prikaza regresije (slika 4.) i analize rezidualnih vrijednosti (slika 5.), procijenite reprezentativnost i pouzdanost postavljenog modela.
h) Na što je potrebno obratiti posebnu pozornost prije računanja nagiba i odsječka, a povezano je s tabličnim podatcima i linearizirani oblicima kemijskih jednadžbi?

2. KLASTERSKA ANALIZA

Suvremene analitičke metode omogućuju prikupljanje velikog broja podataka. U kliničkoj kemiji primjerice postoje brojni parametri koji se svakodnevno analiziraju u krvi, urinu ili kosi, dok kromatografske i spektroskopske metode daju analitičke podatke mnogih komponenata u nekom uzorku. U slučajevima u kojima se mjeri velik broj varijabli, osnovna statistika nije dovoljna, stoga podatci zahtijevaju multivarijantni pristup obradi. Kemijske reakcije mogu se izvoditi pod različitim uvjetima kao što su promjene temperature, tlaka, koncentracije reaktanata, pH-vrijednosti, vrste otapala i sl. Svaka od tih varijabli može utjecati na brzinu i prinos reakcije te se klasterska analiza (engl. *cluster analysis*, CA) u takvim slučajevima može koristiti kao metoda za prepoznavanje sličnosti unutar takvih sustava. Klasterska analiza omogućava grupiranje reakcija koje daju slične rezultate pri istim uvjetima, olakšavajući optimizaciju procesa. U toksikološkim se analizama klasterska analiza koristi za grupiranje uzoraka iz okoliša prema sličnosti njihovih toksikoloških profila, nakon čega je moguće otkriti onečišćena područja i moguće izvore onečišćenja. Nadalje, metoda omogućava da se nove ili nepoznate kemikalije grupiraju prema njihovoj sličnosti što pomaže u predviđanju njihove toksičnosti i pravovremenom otkrivanju moguće opasnih spojeva.

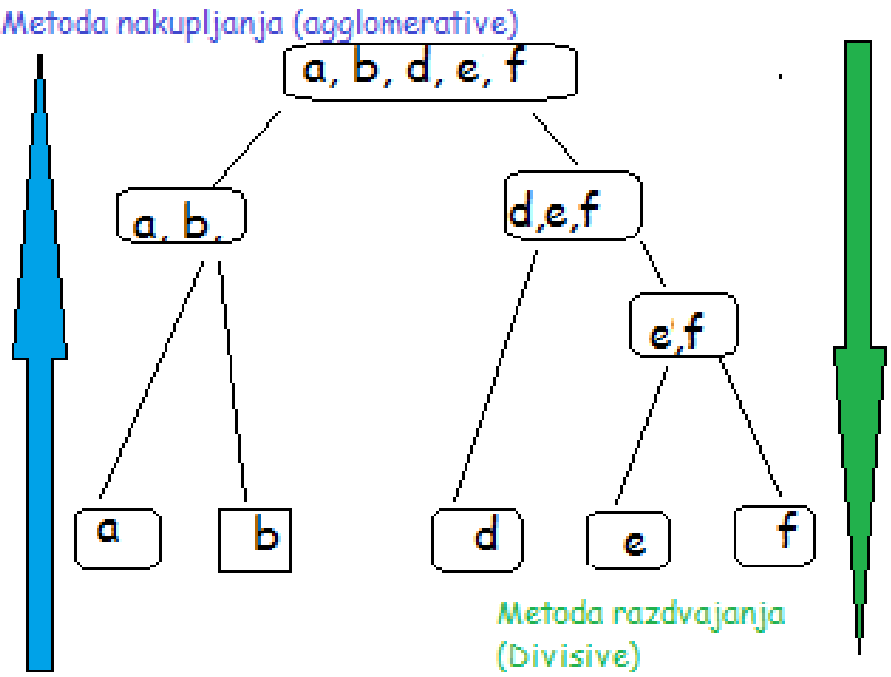
Klasterska je analiza multivarijantna metoda grupiranja objekata u skupine (klastere) na temelju njihove sličnosti tako da su objekti unutar istog klastera što sličniji, a različiti od objekata u drugim klasterima. Metoda je fleksibilna i primjenjiva na različite vrste podataka, a koristi se u prirodnim (kemija, biologija, medicina i sl.) i u društvenim znanostima (psihologija, ekonomija, sociologija i sl.). Ovisno o načinu formiranja klastera, klasterska analiza može biti hijerarhijska i nehijerarhijska.

Hijerarhijska klasterska analiza vrsta je klasterske analize koja stvara hijerarhijsku strukturu klastera (slika 6.), a krajnji je rezultat struktura koja se prikazuje u obliku dendrograma (slika 7.).

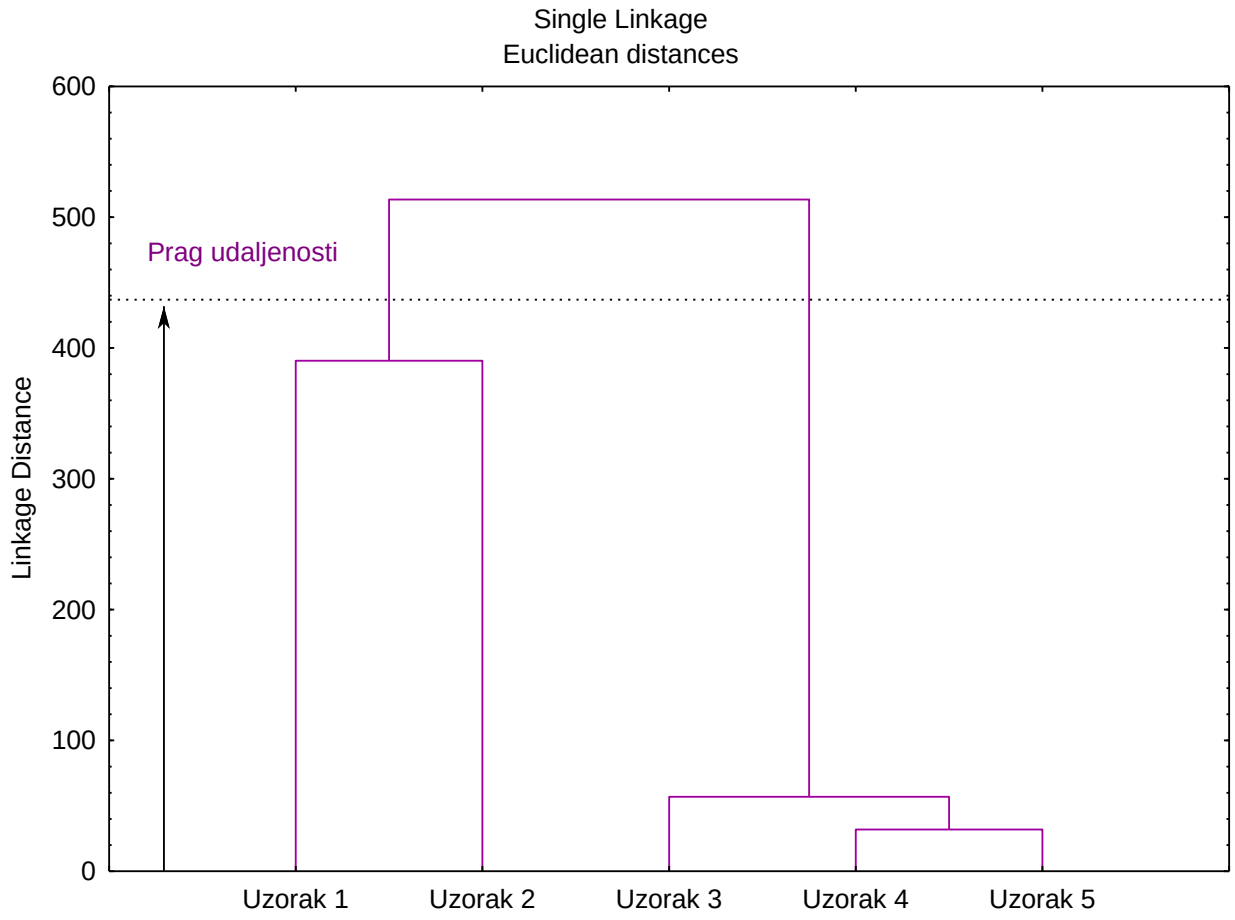
Dva su osnovna pristupa u hijerarhijskoj klasterskoj analizi:

a) metoda nakupljanja (engl. *agglomerative*) u kojoj na početku svaki objekt čini zaseban klaster, nakon čega se slični objekti spajaju u zasebne klastere (slika 7.)

b) metoda razdvajanja (engl. *divisive*) u kojoj na početku svi objekti čine jedan klaster koji se potom dijeli na manje klastere.



Slika 6. Hijerarhijska klasterska analiza



Slika 7. Rezultat klasterske analize. Prikazani su dendrogram i prag udaljenosti. Korištene su metoda najbližih susjeda i euklidska udaljenost.

Na slici 7. prikazan je primjer dendrograma (konačan rezultat klasterske analize) i prag udaljenosti što je granica s pomoću koje se „siječe“ dendrogram kako bi se odredio konačan broj klastera.

Nehijerarhijska klasterska analiza ne koristi stablo kao način prikaza, već izravno grupira podatke u određen broj klastera. Najpoznatija je nehijerarhijska metoda *K*-means algoritam kod kojeg se prije obrade podataka unaprijed određuje broj klastera (*K*). Metoda je jednostavna, brza i pogodna za velik broj rezultata, a nedostatak joj je visoka osjetljivost na izdvojene vrijednosti (engl. *outliers*) koje mogu pomaknuti središte klastera i utjecati na konačne rezultate.

2.1. Osnovni koraci u klasterskoj analizi

Klasterska se analiza temelji na metodološkom pristupu u četirima značajnim koracima:

1. odabir ključnih varijabli

U prvom koraku utvrđuju se i odabiru varijable koje najbolje opisuju obilježja promatranih objekata i omogućuju njihovo grupiranje.

2. odabir mjera sličnosti/udaljenosti

U obradi numeričkih podataka stupanj sličnosti među objektima određuje se izračunom njihove međusobne udaljenosti. Najčešće se koriste mjere udaljenosti prikazane izrazima (13 – 16), ovisno o specifičnosti skupa podataka:

a) euklidska:

$$d=\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \qquad (13)$$

b) kvadrirana euklidska:

$$d = \sum_{i=1}^{k}(x_i - y_i)^2 \qquad (14)$$

c) Manhattan:

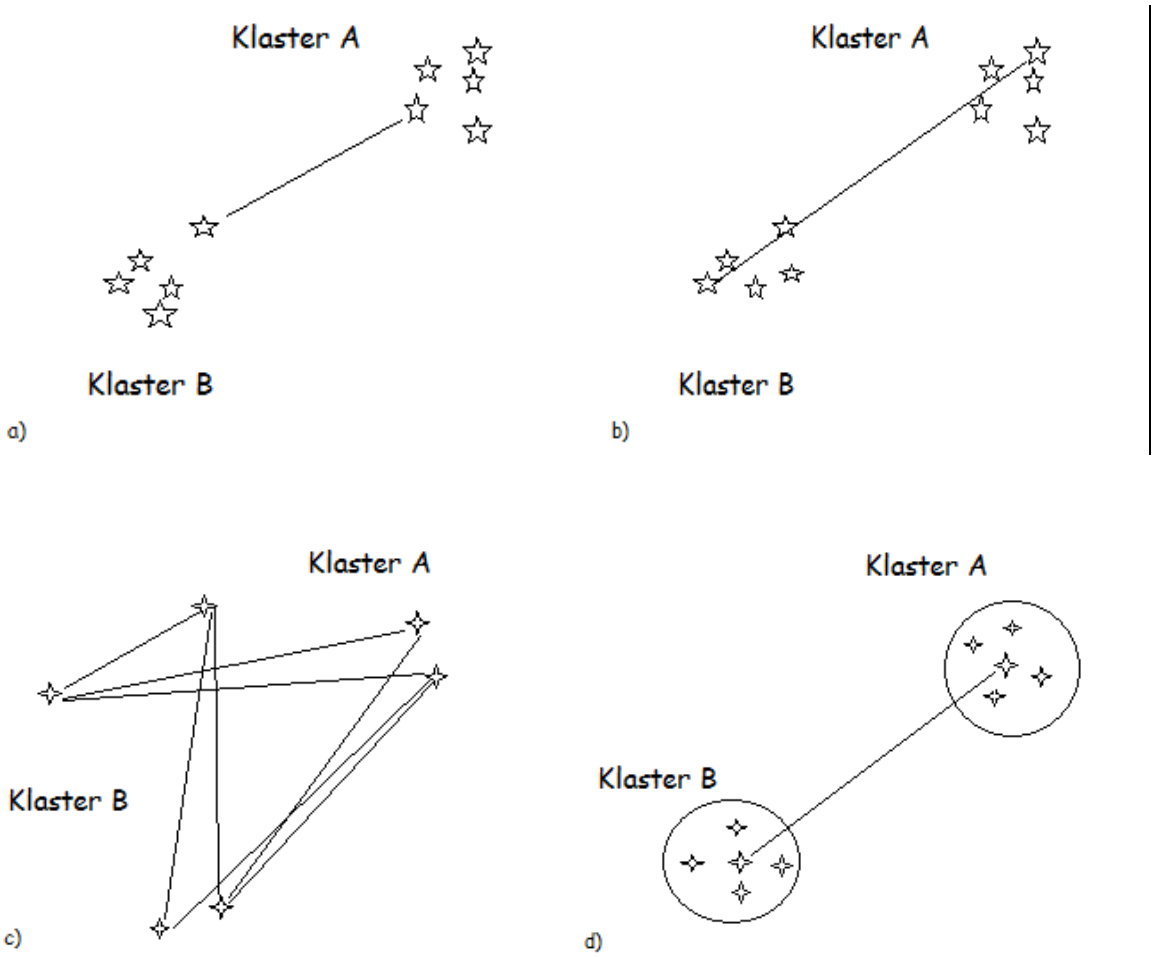
$$d = \sum_{i=1}^{k}|x_i - y_i| \qquad (15)$$

d) Minkowski:

$$d = \left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q} \qquad (16).$$

3. Odabir metode spajanja klastera

Najznačajnije metode, prikazane na slici 8., jesu: metoda najbližih susjeda (engl. *Single Linkage*), metoda najdaljeg susjeda (engl. *Complete Linkage*), metoda prosječne udaljenosti (engl. *Average Linkage*) i metoda centroida (engl. *Centroid*).



Slika 8. Metode spajanja klastera: a) metoda najbližih susjeda (engl. *Single Linkage)*, b) metoda najdaljeg susjeda (engl. *Complete Linkage*), c) metoda prosječne udaljenosti (engl. *Average Linkage*), d) metoda centroida

4. Određivanje broja klastera (interpretacija dendrograma)

Optimalan broj klastera u hijerarhijskoj analizi procjenjuje se povlačenjem horizontalne linije, odnosno „reza“, na određenoj razini udaljenosti unutar dendrograma. Broj vertikalnih linija koje taj rez presijecaju upućuje na broj identificiranih skupina što omogućuje subjektivnu, ali

znanstveno utemeljenu procjenu strukture podataka. Ovisno o postavkama grafičkog prikaza, dendrogram može biti usmjeren vertikalno ili horizontalno što utječe na raspored varijabli na osima koordinatnog sustava.

U vertikalnom prikazu na apscisi se nalaze pojedinačni uzorci ili već formirani klasteri, dok ordinata služi za očitavanje mjere njihove udaljenosti ili sličnosti. Nasuprot tomu, ako je dendrogram prikazan horizontalno (slika 9.), uzorci su poredani duž osi *y*, dok se na osi *x* prikazuju udaljenosti na kojima dolazi do spajanja elemenata. Pravilno razumijevanje te orijentacije presudno je za ispravno očitavanje rezultata prikazanih na slici 7. kao i za daljnju kemometričku interpretaciju analiziranog skupa podataka.

2.2. **Zadatci**

2.2.1. Na temelju podataka u tablici 10., koja prikazuje vrijednosti intenziteta fluorescencije za 12 različitih spojeva pri četirima valnim duljinama: 300 nm, 340 nm, 400 nm i 460 nm, izračunajte euklidsku udaljenost između uzoraka E i F, a potom između uzoraka C i F.

Tablica 10. Vrijednosti intenziteta za 12 različitih spojeva pri četirima valnim duljinama

SPOJ	300 nm	340 nm	400 nm	460 nm
A	16,00	62,01	67,02	27,01
B	15,02	60,03	69,01	31,03
C	14,03	59,01	68,00	31,01
D	15,00	61,02	71,10	31,01
E	14,03	60,02	70,01	30,01
F	14,03	59,01	69,00	30,01
G	17,02	61,03	68,05	28,01
H	16,01	60,00	69,01	28,02
I	15,01	59,02	72,00	30,03
J	17,02	63,01	69,00	27,02
K	18,02	62,02	68,02	28,00
L	18,01	60,01	67,01	28,00

Udaljenost između dviju točaka u *n*-dimenzijskom prostoru s koordinatama $x = (x_1, x_2...x_n)$ i $y = (y_1, y_2..y_n)$ najčešće se izračunava kao euklidska udaljenost. Ona predstavlja geometrijsku

udaljenost „zračnom linijom“ između dvaju objekata, a matematički se definira izrazom (17):

$$d=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_n-y_n)^2} \qquad (17).$$

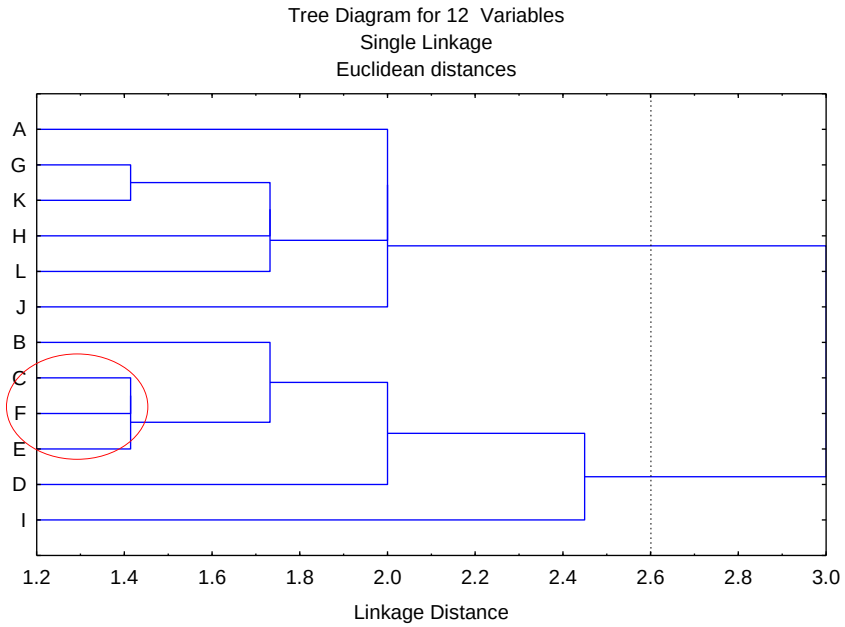
Prema izrazu (17) računa se udaljenost između uzorka E i F:

$d=\sqrt{(14{,}03-14{,}03)^2+(60{,}02-59{,}01)^2+(70{,}01-69{,}00)^2+(30{,}01-30{,}01)^2}=\sqrt{2}$

Izračunajte udaljenost između uzoraka C i F.

Provedite klastersku analizu na temelju podatka prikazanih u tablici 10.

Rezultat provedene klasterske analize prikazan je na slici 9.



Slika 9. Rezultat klasterske analize prema podatcima iz tablice 10. (program *Statistica*)

Pitanja:

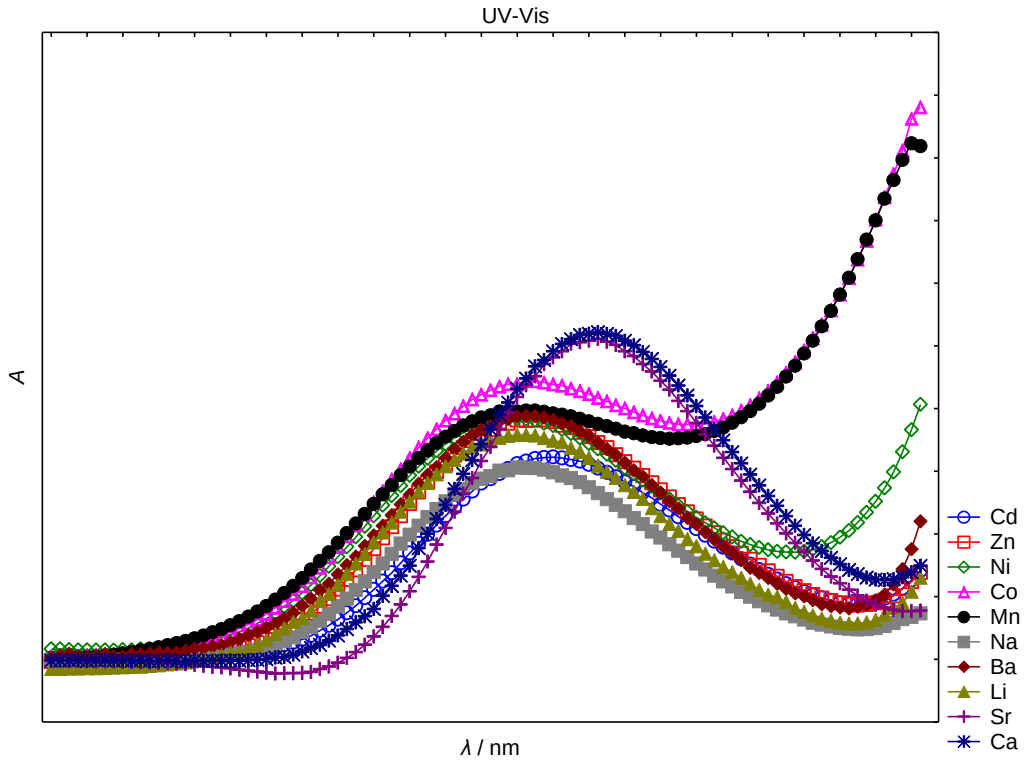
a) Komentirajte rezultat klasterske analize prikazan u obliku dendrograma na slici 9.

b) Ako se upotrijebi neka druga mjera udaljenosti ili druga metoda povezivanja klastera, može li doći do promjena u konačnom rezultatu?

c) Što možete zaključiti o klasteru koji čine uzorci C, F i E?

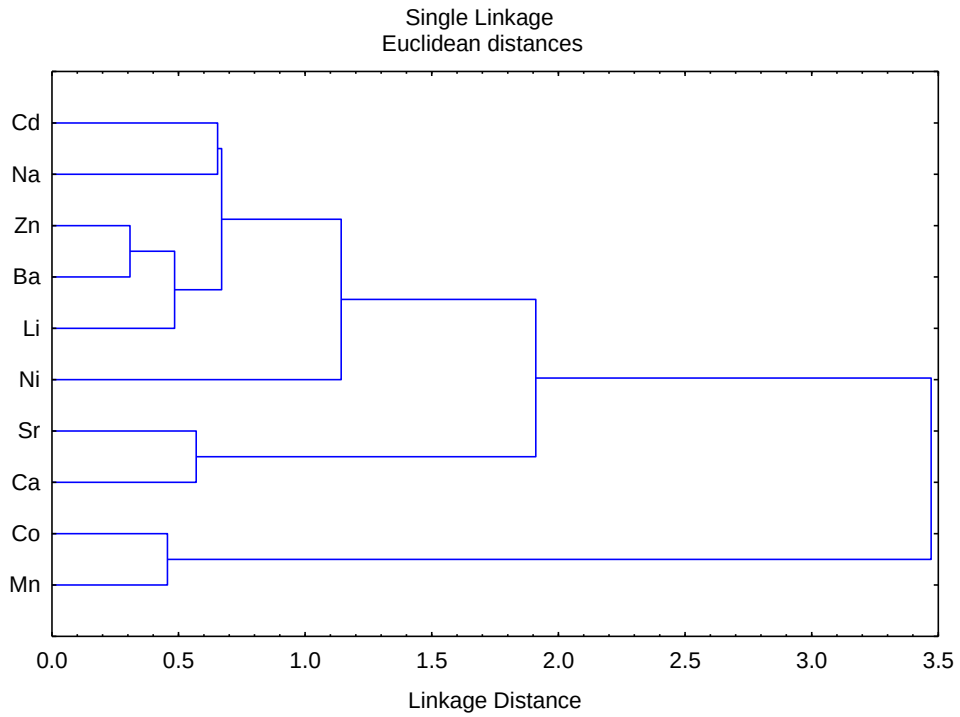
2.2.2. Metodom UV-Vis spektrofotometrije snimljeni su spektri vodenih otopina nitratnih soli različitih elemenata (Cd^{2+}, Zn^{2+}, Ni^{2+}, $Co^{2+,}$ Mn^{2+}, Na^{+}, Ba^{2+}, Li^{+}, $Sr^{2+,}$ Ca^{2+}). Potrebno je utvrditi postoji li pravilnost u grupiranju spektara s obzirom na skupine navedenih elemenata u periodnom sustavu.

Napomena: Budući da vizualnim pregledom spektara nije moguće utvrditi postojanje sličnosti spektara (slika 10.), na rezultate mjerenja potrebno je primijeniti metodu klasterske analize.



Slika 10. UV-Vis spektri vodenih otopina nitratnih soli odabranih elemenata (Cd, Zn, Ni, Co, Mn, Na, Ba, Li, Sr, Ca)

UV-Vis spektri vodenih otopina različitih nitratnih soli pokazuju složene krivulje (slika 10.) zbog čega je vizualno teško donositi zaključke o sličnostima i razlikama među kationima. Primjenom klasterske analize dobiva se objektivna klasifikacija spektara: spektri Ca^{2+} i Sr^{2+} svrstani su u jedan klaster, dok su Co^{2+} i Mn^{2+} stvorili zaseban klaster što u ovom slučaju upućuje na njihovu sličnost (slika 11.).



Slika 11. Rezultat klasterske analize na UV-Vis spektrima (program *Statistica*)

Pitanja:

a) Objasnite šestočlani klaster koji čine Li^{+}, Ba^{2+}, Zn^{2+}, Na^{+}, Cd^{2+}.

b) Napišite elektronske konfiguracije $Ni^{2+,}$ Co^{2+} i Mn $^{2+}$ iona i objasnite zbog čega se nikal izdvojio u zaseban klaster, smješten unutar kompleksnog šestočlanog klastera.

c) Postoje li slični primjeri u kemiji koji bi se mogli riješiti metodom klasterske analize kako bi se dobio bolji uvid u strukturu podataka?

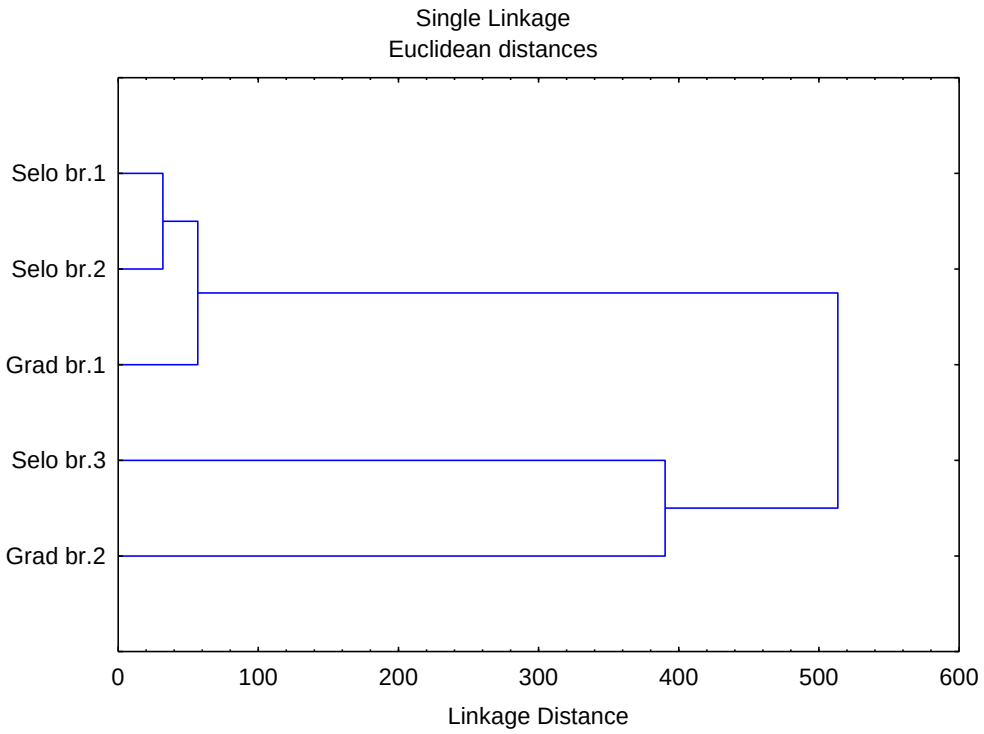
2.2.3. Koncentracije arsena u uzorcima vode za piće određene su primjenom induktivno spregnute plazme uz masenu spektrometriju (engl. *Inductively Coupled Plasma Mass Spectrometry*, ICP-MS), a dobiveni analitički podatci prikazani su u tablici 11. Kako bi se ispitao utjecaj lokacije na kemijski sastav vode, uzorkovanje je provođeno na pet različitih lokacija.

Tablica 11. Koncentracije arsena u vodi na trima ruralnim i dvjema urbanim lokacijama (μg/L)

Selo br. 1	Selo br. 2	Selo br. 3	Grad br. 1	Grad br. 2
0,625	3,739	160,4	0,063	291,5
0,384	0,737	187,9	0,044	37,61
0,833	2,599	197,6	0,59	130,35
0,446	0,521	194,8	0,049	21,94
0,99	4,944	177,1	0,095	123,34
16,91	0,622	200,8	0,081	225,01
2,263	0,409	188,7	0,052	129,3
1,264	8,85	224,2	0,078	127,56
16,168	0,261	30,6	0,067	15,89
0,33	0,145	182,6	56	216,77
8,643	0,281	250,4	0,034	127,23
18,874	0,347	217	0,056	133,55
1,398	0,402	230	0,045	34,45

Rezultat klasterske analize prikazan je na slici 12. Prvu skupinu čine lokacije Grad br. 2 i Selo br. 3 koje su zbog svoje neposredne blizine (manje od 2 km udaljene) pogodne za ispitivanje lokalnih varijacija u istoj vodonosnoj zoni. Nasuprot njima, lokacije Selo br. 1 i Selo br. 2, koje čine zaseban prostorni klaster, smještene su više od 50 km udaljenosti od prve skupine što omogućuje usporedbu regionalnih razlika u kvaliteti vode. Grad br. 1 od ostalih je mjernih postaja udaljen približno 100 km. Takav eksperimentalni dizajn daje osnovu za daljnju kemometričku obradu podataka i otkrivanje mogućih izvora onečišćenja.

Klasterska analiza koncentracija arsena u vodi na pet odabranih lokacija pokazala je da su dvije bliske lokacije (< 2 km) svrstane u jedan klaster, dok su tri udaljene lokacije (> 50 km) tvorile zaseban tročlani klaster. Takva je prostorna raspodjela u skladu s općim obrascem u Slavoniji i Baranji gdje su istočniji dijelovi određeni višim koncentracijama arsena u vodi u odnosu na lokacije smještene u primjerice Baranji i zapadnijim područjima.



Slika 12. Rezultat klasterske analize na koncentracijama arsena u vodi (program *Statistica*)

Pitanja:

a) Objasnite na što upućuje rezultat klasterske analize (s obzirom na različite lokacije), s posebnim osvrtom na lokaciju Grad br. 1.
b) Utječu li regionalne geokemijske razlike na sličnost koncentracija arsena u vodi?
c) Ako se upotrijebi neka druga mjera udaljenosti ili druga metoda povezivanja klastera, hoće li doći do promjene u konačnom rezultatu?

2.2.4. U tablici 12. prikazani su kemijski uzorci i mjerene vrijednosti sljedećih parametara: pH-vrijednost, temperatura i koncentracija kadmija u vodi. Na temelju podataka o pH-vrijednosti, temperaturi i koncentraciji tvari u uzorcima vode, potrebno je provesti klastersku analizu u cilju grupiranja uzoraka prema njihovim sličnostima.

Tablica 12. Izmjerene vrijednosti parametara: pH-vrijednosti, temperature i koncentracija kadmija u vodi

Uzorak	pH	Temperatura (˚C)	Koncentracija (µg/L)
1	7,1	24,9	10,01
2	6,8	21,9	12,02
3	7,1	23,9	11,01
4	5,5	30	25,00
5	5,2	31,9	27,02
6	5,7	31	24,01
7	8,1	19,9	5,02

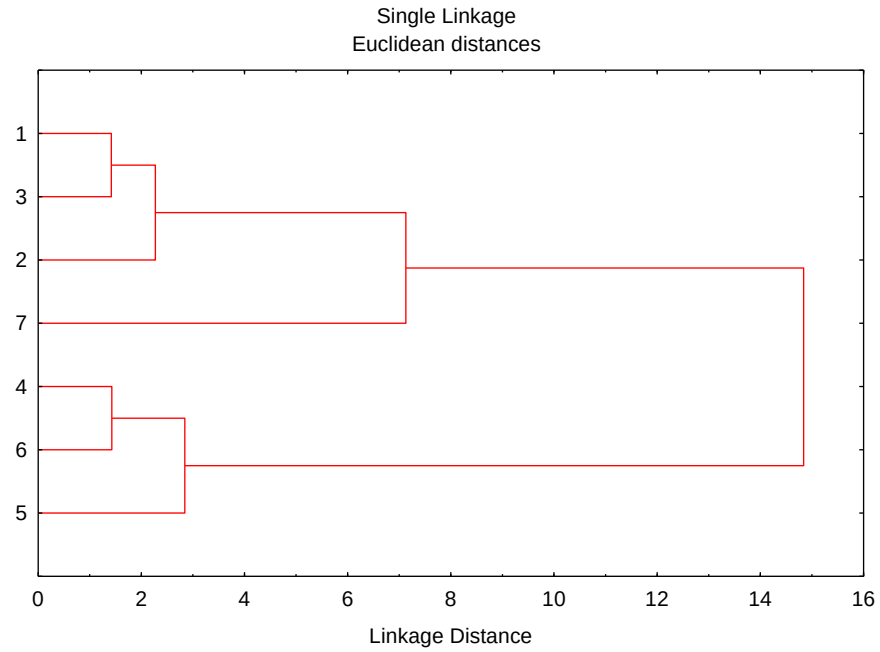
Rezultat klasterske analize uz upotrebu metode najbližih susjeda (engl. *Single linkage*) i euklidske udaljenosti (engl. *Euclidean distance*) prikazan je na slici 13. Upotrijebite neki drugi način povezivanja, odnosno mjeru udaljenosti i provjerite dobije li se isti rezultat.

Klaster koji čine uzorci 1, 3 i 2 predstavljaju uzorke neutralne pH-vrijednosti, umjerene temperature i srednje koncentracije kadmija.

Klaster koji čine uzorci 4, 6 i 5 kombinacija je uzoraka niske pH-vrijednosti, visoke temperature i visoke koncentracije kadmija što može upućivati na povećano onečišćenje i nepovoljne okolišne uvjete. Jednočlani klaster koji čini izdvojeni uzorak 7 značajno je drugačiji od ostalih, a odlikuje se visokom pH-vrijednošću, niskom temperaturom i koncentracijom kadmija u vodi.

Najveće razlike između klastera uočavaju se u koncentraciji kadmija i temperaturi, dok pH-vrijednost dodatno doprinosi jasnom razdvajanju skupina. Klaster koji čine uzorci 4, 6 i 5 značajno odstupa od ostalih zbog viših vrijednosti svih promatranih varijabli.

Zaključak: Posebna pozornost trebala bi se posvetiti uzorcima iz klastera koji čine uzorci 4, 5 i 6 zbog mogućeg negativnog utjecaja na ekosustav.



Slika 13. Rezultat klasterske analize na podatcima pH-vrijednosti, temperature i koncentracija kadmija u vodi (program *Statistica*)

2.2.5. Na temelju relativnih intenziteta karakterističnih apsorpcijskih vrpci dobivenih infracrvenom spektroskopijom s Fourierovom transformacijom (engl. *Fourier Transform Infrared Spectroscopy*, FT- IR) (tablica 13.), potrebno je provesti klastersku analizu uzoraka organskih spojeva, opisati dobivene klastere i povezati ih s prisutnim funkcijskim skupinama.

Tablica 13. Relativni intenziteti karakterističnih apsorpcijskih vrpci u FT-IR spektrima analiziranih organskih spojeva

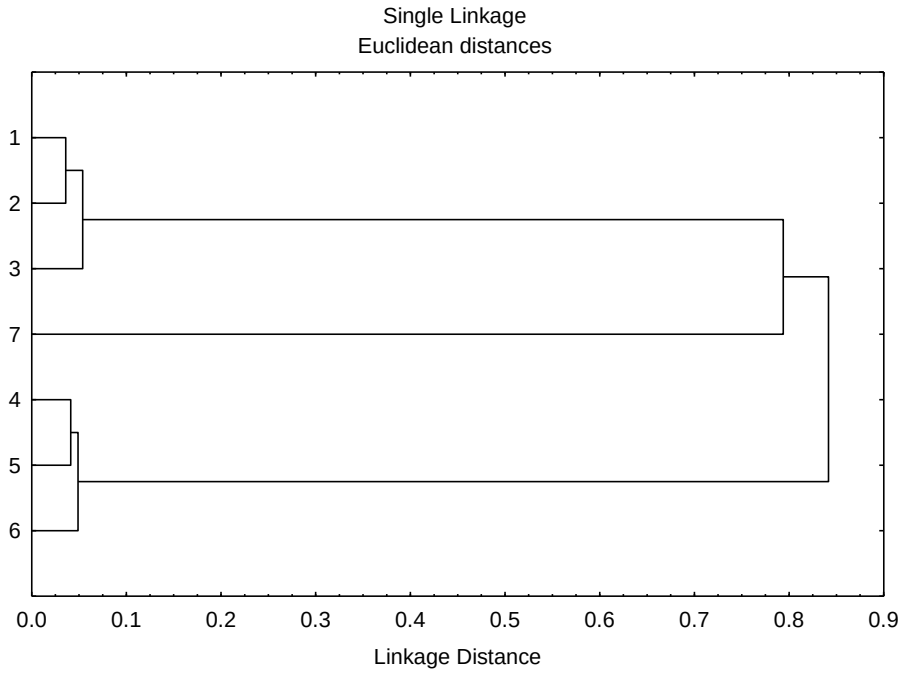
Uzorak	C-N (1300 cm^{-1})	O–H (3220 cm^{-1})	C–H (2875 cm^{-1})
1	0,851	0,121	0,60
2	0,822	0,121	0,58
3	0,881	0,080	0,622
4	0,141	0,901	0,55
5	0,121	0,881	0,52
6	0,180	0,921	0,57
7	0,050	0,051	0,40

Analizom udaljenosti među uzorcima uočeno je razvrstavanje uzoraka u tri klastera što omogućava jasno razdvajanje spojeva prema funkcionalnim skupinama (slika 14.):

1. klaster 1., spojevi koji sadržavaju C-N skupinu, kojemu pripadaju uzorci 1, 2 i 3 (visok intenzitet C-N vrpce, nizak intenzitet O-H vrpce, umjeren intenzitet C-H vrpce)

2. klaster 2., spojevi koji sadržavaju O-H skupinu, kojemu pripadaju uzorci 4, 5 i 6

3. klaster 3., kojemu pripada uzorak 7, sa slabim ili odsutnim C-N, C-H i O-H signalima što upućuje na spojeve bez izraženih polarnih funkcionalnih skupina.



Slika 14. Rezultat klasterske analize na uzorcima apsorpcijskih vrpci FT-IR spektara (program *Statistica*)

Pitanja:

a) Objasnite gdje se u kemijskim istraživanjima primjenjuje klasterska analiza i koje se vrste podataka mogu analizirati tom metodom.
b) Na koji način klasterska analiza pomaže u interpretaciji podataka spektroskopskih mjerenja?
c) Kako se klasterska analiza koristi u toksikološkim analizama?
d) Je li za provođenje klasterske analize neophodna normalna raspodjela podataka?

e) Zbog čega je važno pri izvođenju klasterske analize primijeniti različite tehnike povezivanja i različite mjere udaljenosti?

3. ANALIZA GLAVNIH KOMPONENTI

U kemiji, kao i u ostalim prirodnim i društvenim znanostima, istraživanja često daju opsežne skupove podataka koji obuhvaćaju velik broj mjerenja. Tipični su primjeri u kemijskoj praksi složeni spektralni zapisi ili rezultati multielementnih kemijskih analiza. Takvi podatci često su složeni i teško pregledni, a interpretacija podataka složena, osobito ako su pojedine varijable povezane. U takvim slučajevima primjenjuju se multivarijantne metode obrade podataka.

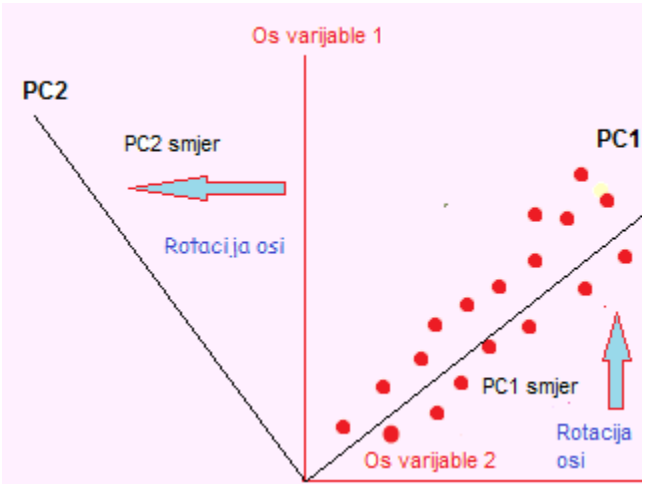
Analiza glavnih komponenenti (engl. *Principal Component Analysis*, PCA) multivarijantna je metoda koja se koristi za smanjenje broja varijabli uz zadržavanje najvažnijih informacija iz podatkovne matrice. PCA pretvara izvorne kemijske varijable u manji broj novih varijabli, tzv. glavne komponente (engl. *Principal Components*, PCs) koje opisuju najveći broj razlika i varijabilnosti u podatcima. Prva glavna komponenta (PC1) objašnjava najveći dio varijabilnosti, a svaka sljedeća komponenta doprinosi objašnjenju preostalog dijela podataka.

U kemijskim istraživanjima PCA nalazi široku primjenu u obradi spektroskopskih podataka, usporedbi kompleksnih uzoraka, identifikaciji sličnosti i razlika među kemijskim sustavima te otkrivanju skrivenih obrazaca. Ta metoda omogućuje učinkovitu vizualizaciju rezultata čime se značajno olakšava donošenje zaključaka na temelju robusnih skupova podataka.

Osim za samu vizualizaciju, PCA se često koristi kao inicijalna faza regresijske analize, poznata kao regresija s glavnim komponentama (engl. *Principal Component Regression*, PCR). U PCR-u se odabrane glavne komponente koriste kao prediktori u regresijskom modelu. Tim se pristupom učinkovito rješava problem multikolinearnosti (međusobne zavisnosti izvornih varijabli) te se značajno poboljšavaju stabilnost, preciznost i pouzdanost dobivenih regresijskih modela. Prema izrazu (18) računa se PC:

$$PC_i = a_{i1} X_1 + a_{i2} X_2 + \cdots a_{ip} X_p \qquad (18)$$

gdje je PC_i – *i*-ta glavna komponenta, X_1, X_2… izvorne su varijable, a a_{ij} predstavlja faktorska opterećenja. Za računanje glavne komponente, izvorna XY-os rotira se tako da odgovara smjeru jediničnog vektora. Nakon rotacije osi, glavne su komponente nove koordinate točaka u odnosu na nove osi (slika 15.).



Slika 15. Pojednostavljen prikaz PCA-a

U nastavku su opisane glavne faze PCA-a, od originalnih podataka do prikaza rezultata.

3.1. Osnovni koraci u analizi glavnih komponenti

a) Odabir i priprema podataka

Uzorke je potrebno odabrati prema kriteriju dostupnosti svih mjerenih varijabli, bez nedostajućih vrijednosti koje bi mogle utjecati na rezultate PCA analize. Također, podatke je potrebno provjeriti u smislu ekstremnih vrijednosti i njihova utjecaja na konačne rezultate.

b) Standardizacija (ako je potrebno)

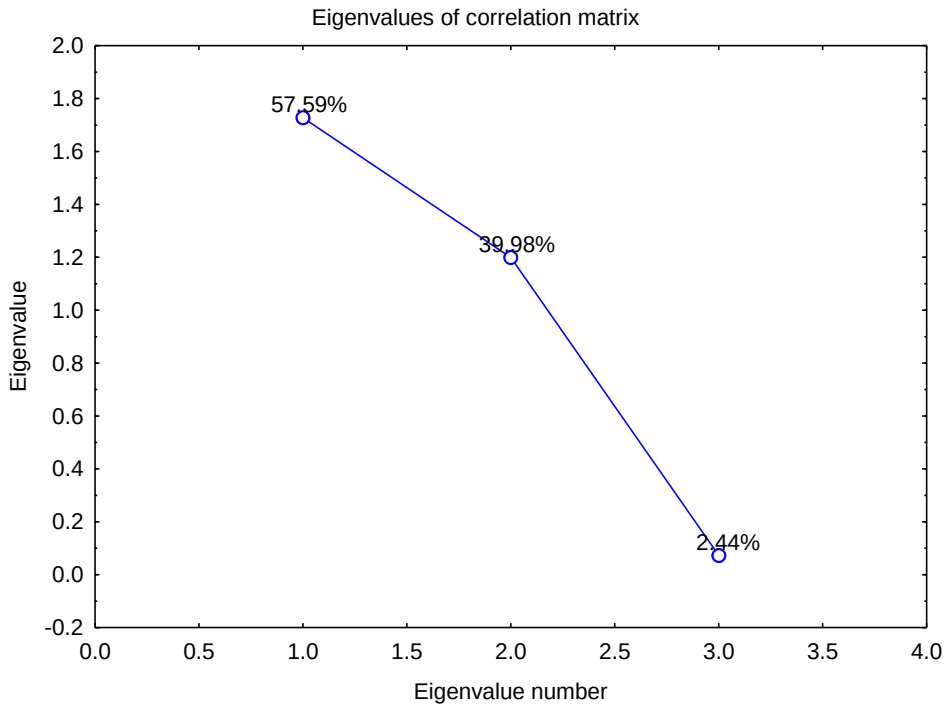
Standardizacija je ključna u PCA analizi jer osigurava da rezultati odražavaju stvarne obrasce u podatcima, a ne razlike nastale uslijed različitih mjernih jedinica (npr. µg/L, %, ˚C) ili raspona varijabli. Standardizacija osigurava jednak doprinos svih varijabli modelu.

c) Izračun glavnih komponenti

Na dobivenoj podatkovnoj matrici provodi se dekompozicija pri čemu se izračunavaju svojstvene vrijednosti i pripadajući svojstveni vektori. Glavne komponente rangirane su prema padajućoj vrijednosti, a u daljnjoj analizi zadržavaju se one komponente koje zajedno objašnjavaju dovoljan udio ukupne varijance podataka.

d) Odabir broja komponenti

Broj glavnih komponenti određuje se na temelju kriterija svojstvenih vrijednosti pri čemu se zadržavaju komponente čije su svojstvene vrijednosti veće od 1 ili analize dijagrama loma (engl. *scree plot*) (slika 16.).



Slika 16. Grafikon svojstvenih vrijednosti. Točka loma predstavlja prelazak s velikog pada na manji pad krivulje. Zadržavaju se točke prije loma jer sadržavaju najveći broj informacija u podatcima (program *Statistica*).

e) Prikaz rezultata PCA

Grafički prikaz rezultata PCA omogućava uvid u strukturu podataka, odnose varijabli, korelacije između varijabli i opažanja kao i moguća grupiranja opažanja.

Dijagram skora koristi se za razumijevanje na koji su način podatci raspoređeni na novim komponentama i može pomoći u prepoznavanju klastera, vrijednosti koje odstupaju ili pri uvidu u strukturu podataka. Dijagram varijabli pokazuje na koji se način varijable odnose prema glavnim komponentama. Bi-plot grafički je prikaz koji kombinira skorove podataka (engl. *scores*) i doprinos varijabli (engl. *variable*) u analizi glavnih komponenata (PCA). Bi-plot omogućava vizualizaciju korelacija odabranih varijabli i skora podataka te tako pomaže u analizi strukture i povezanosti podataka.

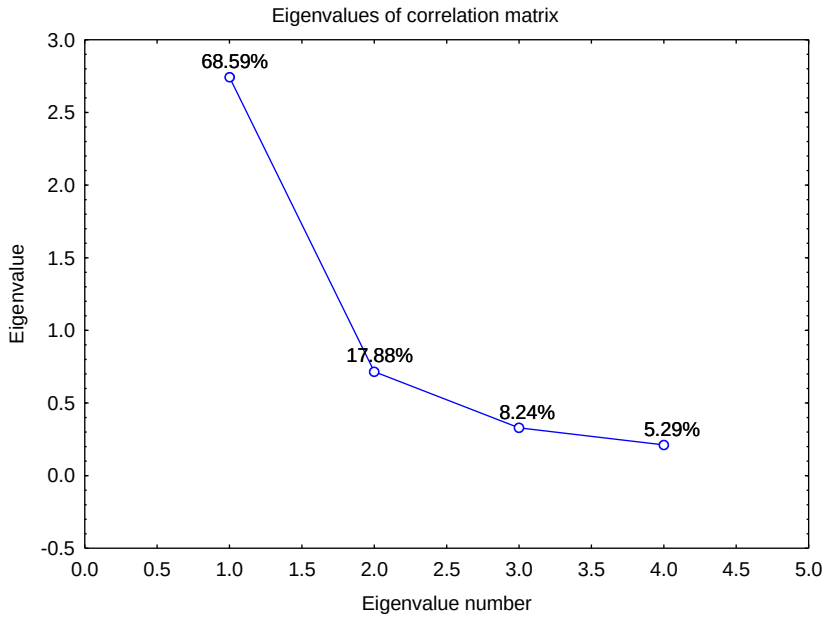
3.2. Zadatci

3.2.1. Primijeniti metodu analize glavnih komponenti na skup podataka prikazan u tablici 10., (podatci se nalaze u poglavlju *Klasterska analiza*) u cilju pronalaženja sličnosti i razlika među uzorcima.

Rezultat PCA prikazan je u tablicama 14. i 15. i slikama 17. i 18.

Tablica 14. Opis rezultata u programu Statistica

Svojstvene vrijednosti (Eigenvalue)	Ukupna varijanca (%) (% Total-variance)	Svojstvene vrijednosti (kumulativno) (Cumulative-Eigenvalue)	Kumulativna varijanca (%) Cumulative -%
1. 2.743620	68.59049	2.743620	68.5905
2.0.715324	17.88311	3.458944	**86.4736**
3. 0.329592	8.23980	3.788536	94.7134
4. 0.211464	5.28661	4.000000	100.0000



Slika 17. Dijagram loma (engl. *scree plot*) (program *Statistica*)

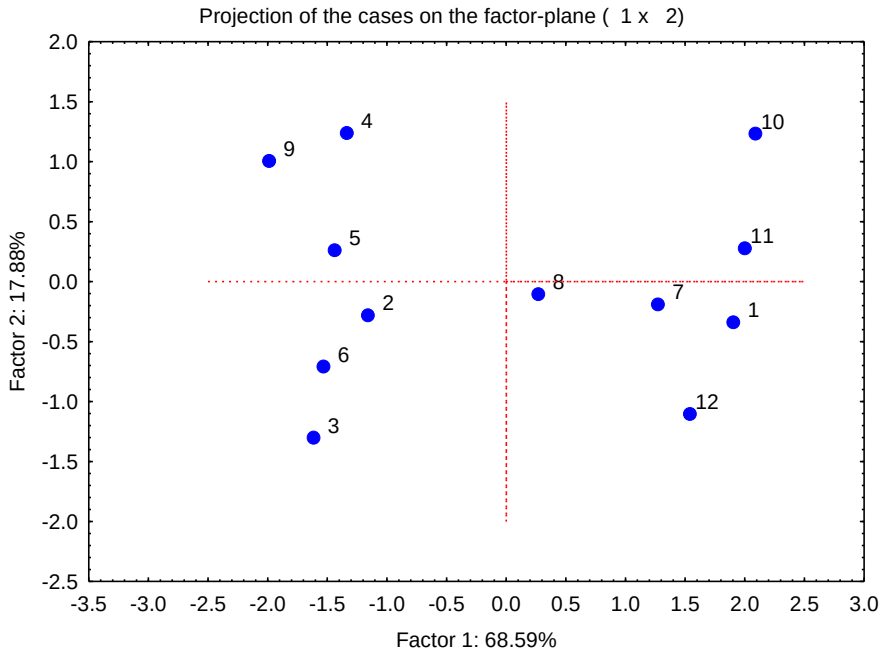
Budući da prve dvije komponente (PC1 i PC2) sadržavaju približno 86 % varijance u podatcima, rezultat PCA moguće je prikazati dvodimenzionalno, odnosno s pomoću dviju komponenti (pratiti tablice 14. i 15. te slike 17. i 18.).

Tablica 15. Vrijednosti faktorskih koordinata (faktorskih opterećenja) uzoraka temeljene na matrici korelacija

Uzorak br.	**Faktor 1**	**Faktor 2**	**Faktor 3***	**Faktor 4***
1	1.90456	-0.33919	-0.838132	0.530710
2	-1.15963	-0.27967	-0.274665	-0.637057
3	-1.61699	-1.30032	-0.465391	-0.204897
4	-1.33697	1.24098	-0.305386	-0.629685
5	-1.43878	0.26139	-0.391765	0.321777
6	-1.53383	-0.70625	-0.132873	0.411336
7	1.27148	-0.18888	0.222376	-0.047771
8	0.26934	-0.10351	0.487993	0.578251
9	-1.98869	1.00577	1.001261	0.359530
10	2.08953	1.23446	-0.419235	0.195481
11	2.00123	0.27876	0.184930	-0.576862
12	1.53874	-1.10353	0.930887	-0.300812

*Vrijednosti faktora 3 i 4 u tablici 15. označene su svjetlijom bojom jer je rezultat PCA moguće prikazati s pomoću prvih dviju komponenti.

PCA analiza vrijednosti fluorescencije snimanih na četirima valnim duljinama pokazuje jasno razdvajanje uzoraka (slika 18.). Uzorci 1, 7, 8, 10, 11 i 12 grupirani su na jednoj strani prostora glavnih komponenti, dok su uzorci 2, 3, 4, 5, 6 i 9 na suprotnoj strani. Takav raspored uzoraka pokazuje da postoje razlike u vrijednostima intenziteta među uzorcima na promatranim valnim duljinama, a prve dvije glavne komponente učinkovito opisuju tu varijabilnost.



Slika 18. Rezultat PCA na vrijednostima intenziteta za 12 različitih spojeva pri četirima valnim duljinama (program *Statistica*)

Pitanja:

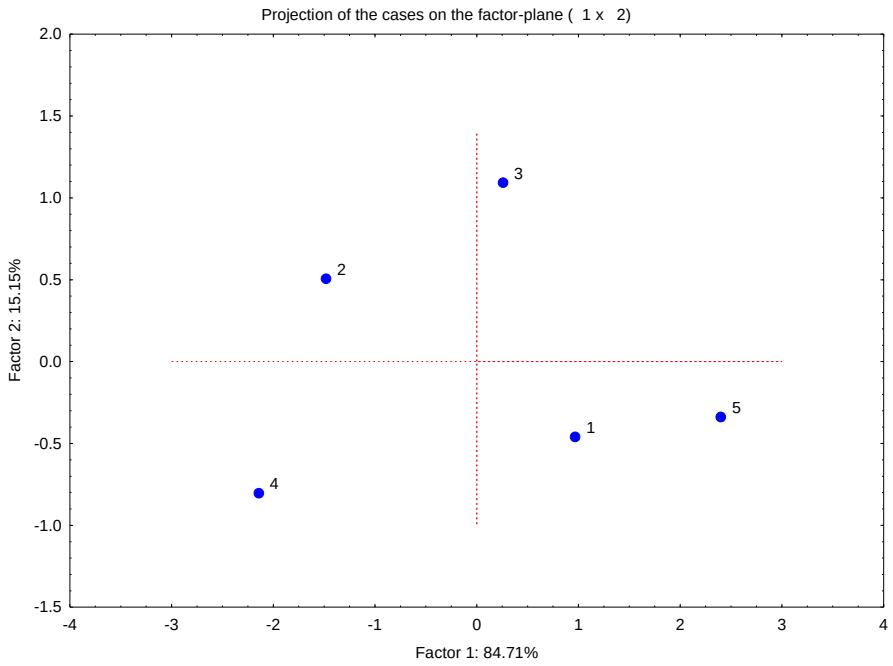
a) Komentirajte poziciju uzorka br. 8.
b) Usporedbom rezultata tablice 15. i slike 18. provjerite jesu li sve točke točno ucrtane u dijagram.
c) Zbog čega su prve dvije komponente u tom primjeru bile dovoljne za opis rezultata?
d) U kojem je slučaju rezultat PCA potrebno prikazati trodimenzionalnim grafičkim prikazom?

3.2.2. U tablici 16. prikazani su rezultati analize četiriju elemenata u uzorcima kose. Metodom PCA potrebno je ispitati postoje li grupiranja među uzorcima.

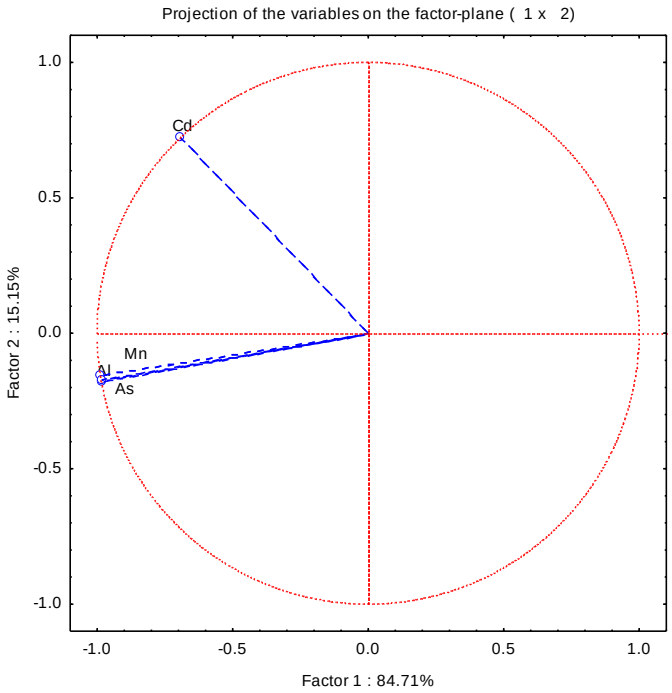
Tablica 16. Rezultati analize četiriju elemenata u uzorcima kose (µg/g)

Uzorak br.	Al	As	Cd	Mn
1	200	300	100	360
2	380	580	420	840
3	200	320	401	380
4	500	760	250	1061
5	50,01	70,03	24,01	100,05

Podatci su obrađeni programom *Statistica,* a rezultati su prikazani na slikama 19. i 20.



Slika 19. Rezultat PCA (opažanja) koncentracija odabranih elemenata u kosi (program *Statistica*)



Slika 20. Rezultat PCA (varijable) koncentracija odabranih elemenata u kosi (program *Statistica*)

Za prikaz rezultata bile su dovoljne prve dvije komponente na što upućuje i dijagram loma. Usporedbom slika 19. i 20. vidljivo je grupiranje uzoraka 1, 5 i 3 kao i uzoraka 2 i 4. Uzorak br. 4 povezan je s višim koncentracijama Al, As i Mn.

Pitanja:

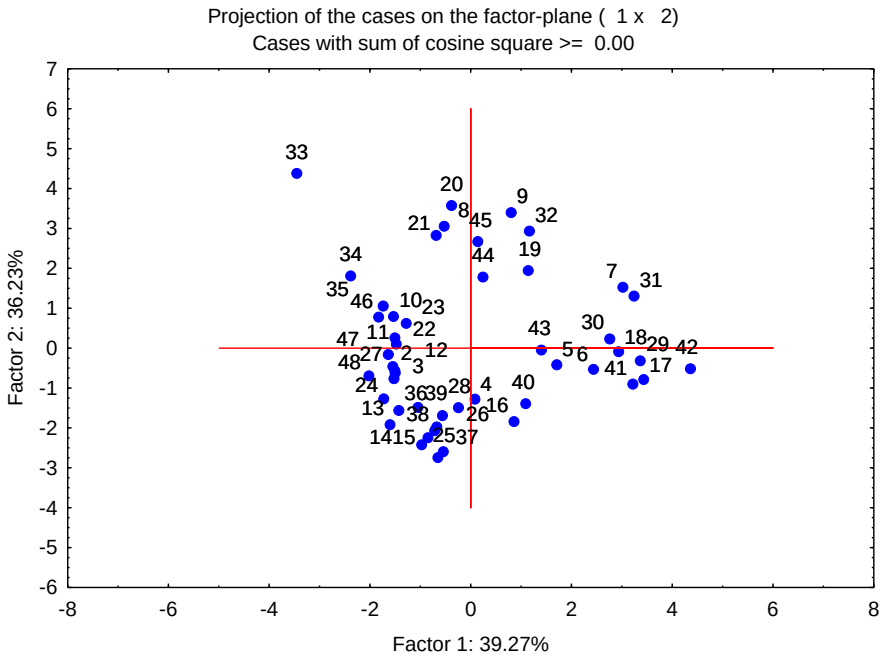
a) Što je moguće zaključiti o uzorcima br. 1, 5 i 3?
b) Komentirajte poziciju uzorka br. 2. S kojim je varijablama visoko koreliran?
c) Zbog čega je rezultat PCA bilo moguće prikazati s pomoću dviju glavnih komponenti?
d) Za razliku od prethodnog primjera, u ovom je primjeru bilo poželjno rezultate PCA prikazati s pomoću dvaju dijagrama. Objasnite zbog čega je tomu tako.

3.2.3. Tablica 17. služi isključivo kao ilustrativni prikaz dijela prikupljenih podataka koncentracija troposferskog ozona (O_3), dušikova (II) oksida (NO) i odabranih meteoroloških parametara (temperatura (T_1, T_2), brzine vjetra (BV), tlaka (P), evaporacije (E), broja požara (F) i količine padalina (R)), dok je cjelokupno analizirano razdoblje obuhvaćalo 48 mjeseci. Prosječni mjesečni podatci varijabli prikazanih u tablici podvrgnuti su metodama PCA i klasterske analize.

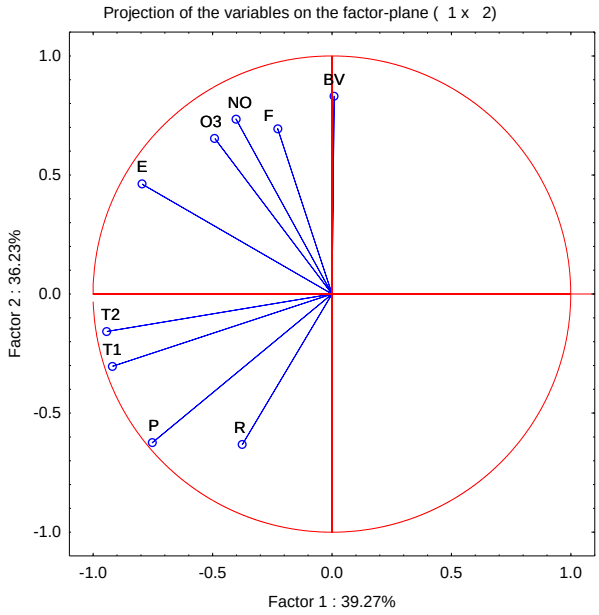
Tablica 17. Izmjerene vrijednosti koncentracija ozona, NO_2, temperatura (maksimalna – T_1, minimalna – T_2), brzine vjetra (BV), tlaka (P), evaporacije (E), broja požara (F) i količine padalina (R) (program *Statistica*).

	1 F	2 R	3 E	4 P	5 T2	6 T1	7 BV	8 O_3	9 NO
1	7	346	123	2.452	29.9	19.8	2.95	254.2133	2.9909771
2	9	40	136.6	2.588	31.6	19.8	2.63	254.4148	3.1714567
3	20	45	139.3	2.577	32.2	19.9	2.81	247.1767	2.92215
4	10	130	117.6	2.217	29.8	19.6	2.72	249.1586	2.7318346
5	10	87	111.2	1.883	28.1	17.8	3.2	246.2533	2.5699825
6	6	50	93.9	1.772	27.6	16.8	2.81	252.4241	2.7046518
7	28	24	110	1.417	26.9	14	3.29	260.7033	2.4319458
8	19	11	144.8	1.652	30	18	3.57	266.9709	3.6068714
9	9	80	131.4	1.601	27.3	15.6	3.79	273.2241	3.7632581

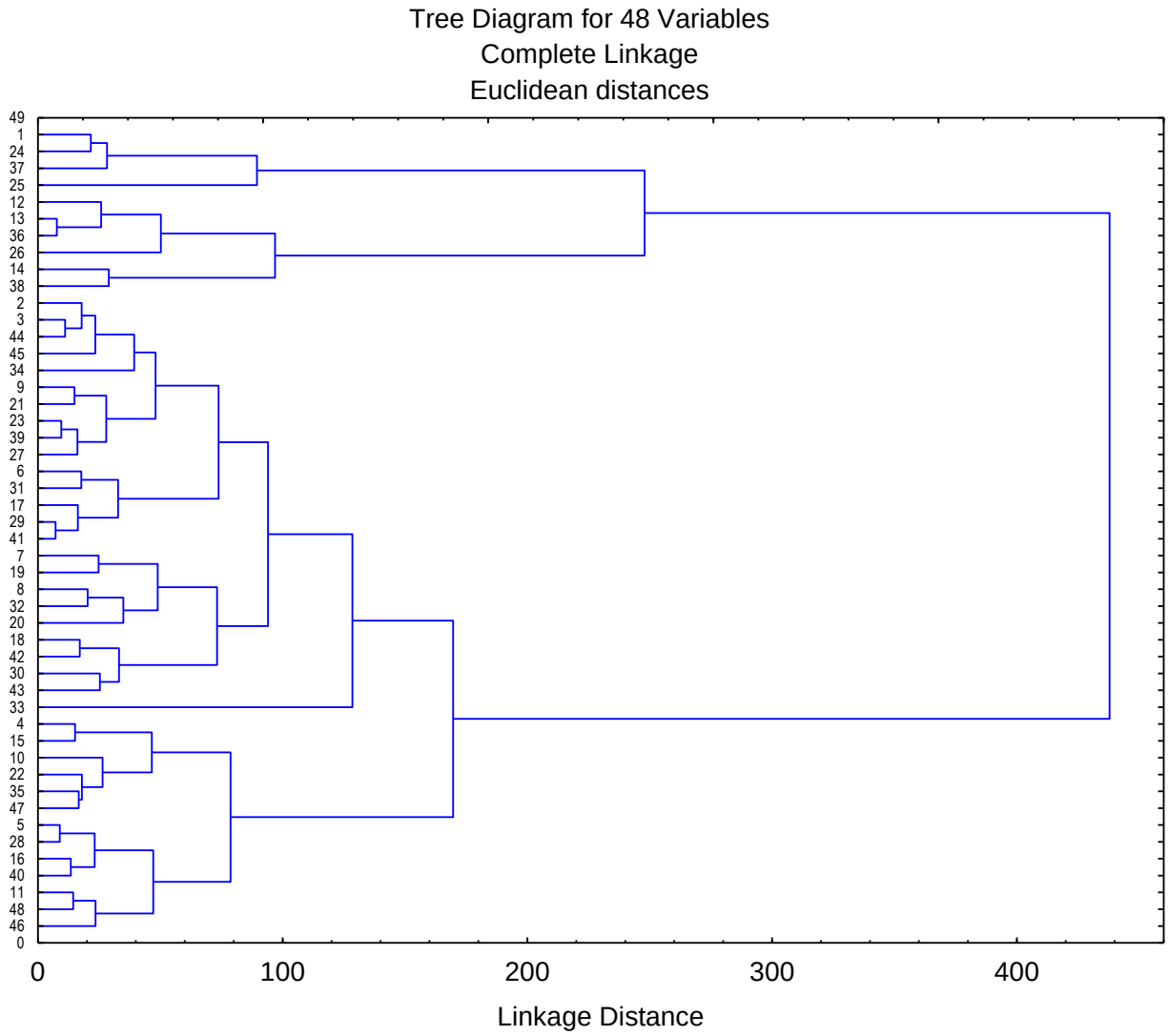
Rezultat PCA prikazan je na slikama 21. i 22., a rezultat klasterske analize na slici 23.



Slika 21. Rezultat PCA na podatcima koncentracija O_3, NO_2, temperatura (T_1, T_2), brzine vjetra (BV), tlaka (P), evaporacije (E), broja požara (F) i količine padalina (R) mjerenih tijekom 48 mjeseci (program *Statistica*)



Slika 22. Rezultat PCA na podatcima: koncentracija O_3, NO, temperatura (T_1, T_2), brzine vjetra (BV), tlaka(P), evaporacije(E), broja požara (F) i količine padalina(R) (program *Statistica*)



Slika 23. Rezultat klasterske analize na podatcima: koncentracija O_3, NO, temperatura (T_1, T_2), brzine vjetra (BV), tlaka (P), evaporacije (E), broja požara (F) i količine padalina (R) (program *Statistica*)

Pitanja:

a) Komentirajte rezultate PCA usporedbom slika 21. i 22.

b) Usporedite rezultate dobivene metodom PCA i metodom klasterske analize.

c) Koje su prednosti metode PCA u odnosu na metodu klasterske analize?

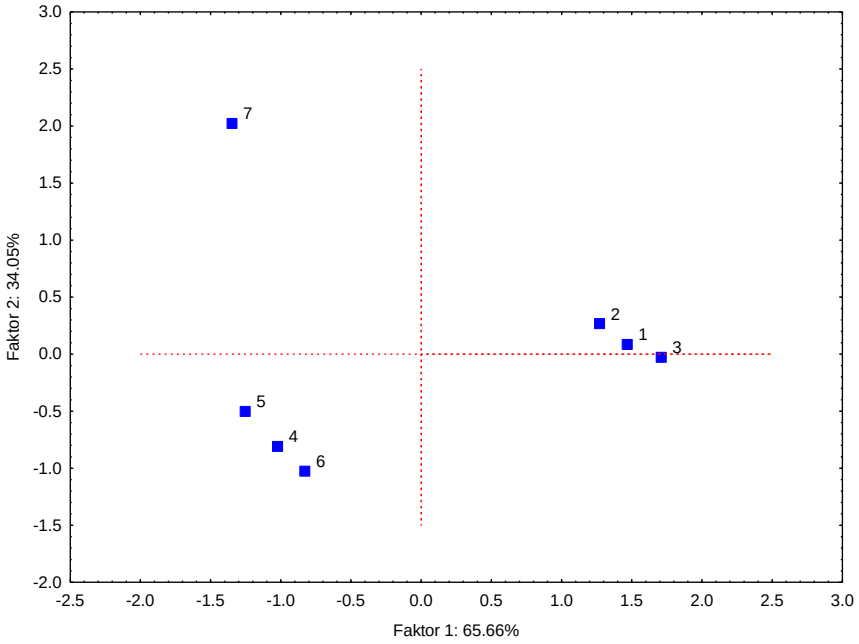
d) Na koji način treba transformirati podatke nakon PCA za metodu klasterske analize?

3.2.4. Na temelju intenziteta odabranih apsorpcijskih vrpci u FT-IR spektrima prikazanim u tablici 18., potrebno je provesti PCA, usporediti rezultate s rezultatom klasterske analize i objasniti prednosti PCA u analizi podataka.

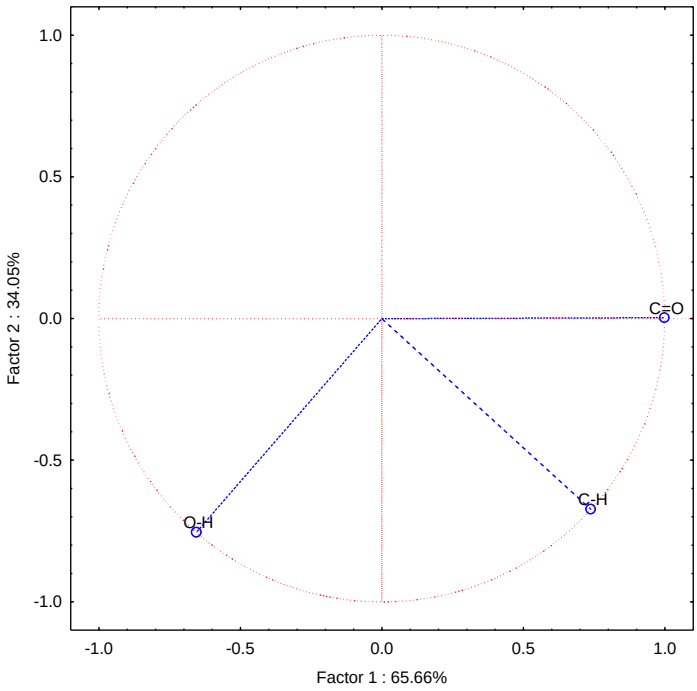
Tablica 18. Relativni intenziteti apsorpcijskih vrpci u FT-IR spektrima

Uzorak	C=O (1630 cm^{-1})	O–H (3220 cm^{-1})	C–H (2875 cm^{-1})
1	0,851	0,121	0,60
2	0,822	0,121	0,581
3	0,881	0,080	0.622
4	0,141	0,901	0,55
5	0,121	0,881	0,52
6	0,180	0,921	0,57
7	0,050	0,051	0,401

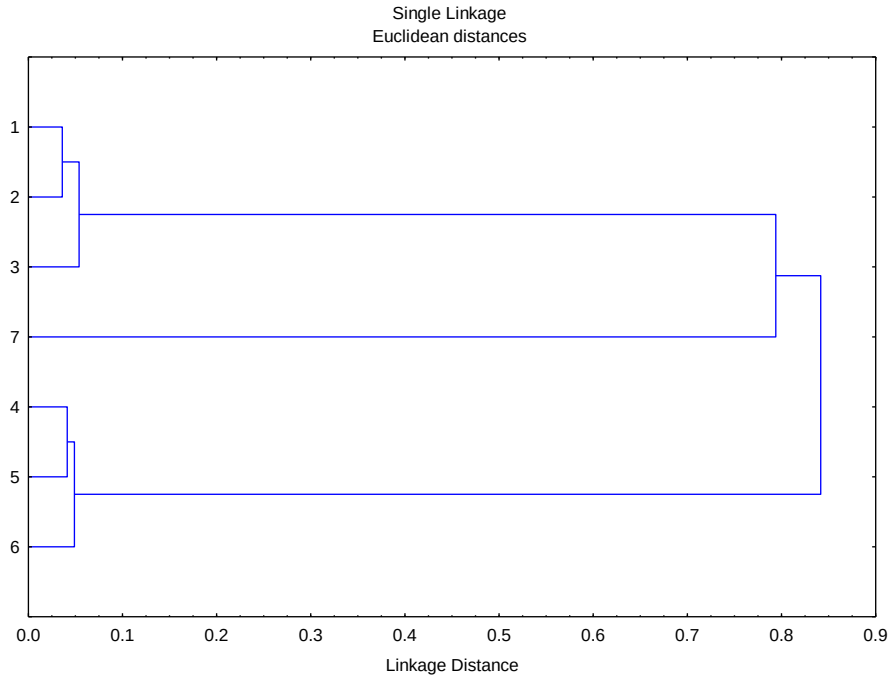
Rezultat PCA prikazan je na slikama 24. i 25., a klasterske analize na slici 26.



Slika 24. Rezultat PCA (uzorci) na podatcima intenziteta apsorpcijskih vrpci FT-IR spektara (program *Statistica*)



Slika 25. Rezultat PCA (varijable) na podatcima intenziteta apsorpcijskih vrpci FT-IR spektara (program *Statistica*)



Slika 26. Rezultat klasterske analize na podatcima intenziteta apsorpcijskih vrpci FT-IR spektara (program *Statistica*)

Primjena PCA na podatke omogućila je jasno razdvajanje uzoraka prema dominantnim funkcionalnim skupinama. Za razliku od metode klasterske analize, PCA je omogućio uvid u varijable koje su najviše doprinijele razdvajanju uzoraka te njihovu identifikaciju.

Pitanja:

a) Koja je osnovna svrha PCA metode?
b) Na koji način PCA smanjuje dimenzionalnost podataka?
c) Koji su osnovni koraci u PCA metodi?
d) Što je svojstvena vrijednost i kakva je njezina uloga u PCA?
e) Na koji se način odabire broj komponenti koje je potrebno zadržati?
f) Zbog čega je i kada važna standardizacija podataka prije izvođenja PCA?
g) Koja su ograničenja PCA, a koja klasterske analize u prethodnom primjeru?
h) Opišite primjene PCA u spektroskopskim mjerenjima.

4. ANALIZA GLAVNIH KOMPONENTI S ROTACIJOM FAKTORA

Iako PCA daje matematički optimalna rješenja u smislu postotka objašnjene varijance u podatcima, dobivene komponente mogu biti kompliciranе za interpretaciju jer pojedine varijable mogu imati značajna opterećenja na većem broju komponenti. U praksi se vrlo često primjenjuje rotacija faktora čime se postiže jednostavnija i lakše objašnjiva struktura faktorskih opterećenja (engl. *factor loadings*). Rotacijom se osi zaokreću u prostoru varijabli bez promjene ukupne objašnjene varijance čime se naglašavaju jaka opterećenja, a „potiskuju" slaba.

Najčešće je korištena *varimax* rotacija, a postoje još i mnoge druge: *quartimax, equamax* itd. Kombinacijom PCA i rotacije faktora omogućava se identifikacija latentnih varijabli koje je moguće povezati s kemijskim, fizikalnim ili drugim prirodnim procesima koji stoje u pozadini izmjerenih podataka. Rotacija faktora ne mijenja postotak objašnjene varijance u podatcima, već isključivo poboljšava mogućnost objašnjenja faktora.

PCA s rotacijom faktora često se koristi u kemiji, posebice u području kemometrike za obradu i interpretaciju složenih skupova podataka. Rotacijom faktora postiže se jasnija struktura faktorskih opterećenja čime se postiže lakše razumijevanje strukture podataka.

U analizi okoliša rotirani faktori mogu dati informacije o mogućem izvoru onečišćenja na temelju analiza pojedinih elemenata ili spojeva u uzorcima vode, tla, biljaka ili zraka.

U području mjerenja metodom nuklearne magnetske rezonancije (NMR) rotacija se faktora primjenjuje kod obrade velikog broja spektara dobivenih iz primjerice bioloških uzoraka ili serije polimera. Nakon rotacije faktora postižu se veća i izraženija opterećenja na kemijski povezanim signalima.

U UV-Vis spektroskopiji, osobito kod smjesa ili serija uzoraka kod kojih se apsorpcijske trake preklapaju, rotacija faktora pomaže u izdvajanju skrivenih komponenti koje doprinose ukupnom spektru.

Općenito, primjena faktorske rotacije u kemiji doprinosi smanjenju dimenzionalnosti podataka, povećanju interpretabilnosti rezultata i boljem razumijevanju složenih kemijskih sustava.

4.1. Zadatci

4.1.1. Primjenom analize glavnih komponenti na podatke iz tablice 19., potrebno je usporediti rezultate dobivene izvornim i rotiranim faktorskim rješenjem. Naglasak je na razumijevanju doprinosa rotacije u jasnijem definiranju strukture podataka i identifikaciji skrivenih uzročnih veza između meteoroloških uvjeta i koncentracija O_3 i NO_2. Tablica 19. obuhvaća dvogodišnji skup mjesečnih prosjeka odabranih fizikalno-kemijskih parametara zraka.

Tablica 19. Koncentracije O_3 (DU), NO_2 i vrijednosti odabranih meteoroloških parametara

Mjesec	Požari	Padaline	Isparavanja	Tlak	T max	T min	Brzina vjetra	O_3	NO_2
1	7	346	123	2,452	29,9	19,8	2,95	254,21	2,99
1	9	40	136,6	2,588	31,6	19,8	2,63	254,41	3,17
1	20	45	139,3	2,577	32,2	19,9	2,81	247,17	2,92
1	10	130	117,6	2,217	29,8	19,6	2,72	249,15	2,73
1	10	87	111,2	1,883	28,1	17,8	3,2	246,25	2,56
1	6	50	93,9	1,772	27,6	16,8	2,81	252,42	2,7
1	28	24	110	1,417	26,9	14	3,29	260,7	2,43
1	19	11	144,8	1,652	30	18	3,57	266,97	3,6
1	9	80	131,4	1,601	27,3	15,6	3,79	273,22	3,76
1	7	148	149,3	2,094	30,2	19,5	3,03	277,97	2,87
1	7	116	144,1	2,24	30,8	20,3	3,08	262,88	3,03
1	3	184	141,2	2,433	31	19,5	3,01	262,85	2,91
2	1	207	129,7	2,576	30,8	20,7	2,3	261,81	2,87
2	2	250	123,5	2,531	31,3	19,3	2,41	252,24	2,7
2	1	140	124,1	2,552	31,3	20,2	2,53	251,27	2,52
2	3	109	111,1	2,153	29,7	18,6	2,6	245,05	2,55
2	6	64	101,1	1,66	26,6	14,4	2,58	240,31	2,8
2	5	30	92,5	1,709	27,4	15,3	2,96	252,39	2,75
2	31	46	116,4	1,578	29,1	16	3,09	252,28	3,6
2	53	15	139,4	1,63	29,9	17,4	3,44	263	3,57
2	11	68	139,2	1,824	29,8	17,6	3,59	271,02	3,77
2	4	150	133,2	2,162	31,2	20,5	2,97	257,41	3,49
2	13	87	127,4	2,263	30,5	20,9	3,13	254,81	3,62
2	25	343	111,6	2,55	31,5	20,7	2,94	253,15	3,16

Rezultat PCA bez rotacije faktora prikazan je u tablici 20.

Tablica 20. Faktorske koordinate varijabli (faktorska opterećenja) (*program Statistica*)

Factor coordinates of the variables, based on correlations		
	Factor 1	**Factor 2**
Požari	-0.490384	-0.341477
Padaline	0.700258	0.017948
Isparavanja	0.254952	-0.882534
Tlak	0.976729	-0.053866
Tmax	0.806156	-0.478643
T min	0.885070	-0.359427
Brzina vjetra	-0.692784	-0.595370
O_3	-0.212613	-0.770798
NO_2	-0.292088	-0.771097

Rezultat PCA s *varimax* rotacijom prikazan je u tablici 21.

Tablica 21. Faktorske koordinate varijabli (faktorska opterećenja) (*program Statistica*)

Factor Loadings (Varimax raw) Extraction: Principal components (Marked loadings are >.700000)		
	Factor - 1	**Factor - 2**
Požari	-0.355342	0.480433
Padaline	0.657702	-0.241065
Isparavanja	0.553933	0.694561
Tlak	0.942616	-0.261489
Tmax	0.916926	0.195535
T min	0.953546	0.057336
Brzina vjetra	-0.465863	0.785739
O3	0.045196	0.798305
NO_2	-0.030005	0.824018

Primjena *varimax* rotacije dovela je do djelomičnog razdvajanja faktora tako da meteorološki podatci dominiraju u prvom, a onečišćivači na drugom pri čemu su i faktorska opterećenja povećana u odnosu na metodu PCA bez rotacije faktora. Primjena *varimax* rotacije otkrila je vezu između brzine vjetra i koncentracija onečišćivača što je važan pokazatelj da je brzina vjetra bitan čimbenik u prijenosu onečišćenja.

Primjena rotacije faktora važna je jer omogućava jasniju interpretaciju rezultata, faktori postaju „čišći“ što znači da svaki faktor dominira nad jednom grupom varijabli (u ovom primjeru: meteorološki, onečišćivači). To omogućava sigurnije povezivanje faktora sa stvarnim procesima u okolišu i razumijevanju uzročno-posljedičnih odnosa. Veća faktorska opterećenja i čišći faktori služe kao pouzdana osnova za daljnje analize poput klasterske analize i regresije s glavnim komponentama.

Pitanja:

1. Kako se interpretacija komponenti promijenila nakon *varimax* rotacije ?
2. Jesu li varijable nakon rotacije jasnije povezane s jednom komponentom?
3. Koje varijable imaju najveća faktorska opterećenja i na koji način to može pomoći u objašnjavanju komponenti?
4. Mogu li se nakon rotacije imenovati komponente i na temelju čega?
5. Po kojim se kriterijima zaključuje da je rješenje s *varimax* rotacijom bolje?
6. Bi li rezultat bio drugačiji da se koristi neka druga vrsta rotacije?

5. REGRESIJA S GLAVNIM KOMPONENTAMA

PCA metoda koristi se kao jedna od faza metode regresija s glavnim komponentama (engl. *Principal Components Regression,* PCR). PCR je metoda koja kombinira dvije metode: analizu glavnih komponenata (PCA) i linearnu regresiju. PCR se koristi za rješavanje problema multikolinearnosti u podatcima, gdje su nezavisne varijable visoko korelirane, što može negativno utjecati na točnost modela u klasičnoj metodi linearne regresije. PCR se koristi metodom PCA kako bi smanjio broj varijabli u regresiji. Umjesto izravnog korištenja originalnih varijabli koje mogu biti visoko korelirane i stvoriti probleme u linearnom modelu, PCR prvo primjenjuje PCA kako bi transformirao podatke u nove varijable, tj. glavne komponente. Nakon te faze, koristi se linearna regresija za modeliranje odnosa između glavnih komponenti i zavisne varijable.

5.1. **Zadatci**

5.1.1. Na temelju danog skupa podataka (tablica 22.), potrebno je analizirati utjecaj više međusobno koreliranih nezavisnih varijabli (A_1, A_2, A_3, A_4, A_5, A_6) na zavisnu varijablu (*c*) primjenom metode regresije s glavnim komponentama (PCR).

Tablica 22. Rezultati mjerenja vrijednosti apsorbancije na različitim valnim duljinama (A_1, A_2 itd..) za deset vrsta i koncentracija jednog od sastojaka (*c*)

Vrsta	*c* (mol/L)	A_1	A_2	A_3	A_4	A_5	A_6
1	0,88	18,71	26,8	42,21	56,61	70,2	83,3
2	0,46	31,31	33,41	45,7	49,31	53,8	55,31
3	0,45	30,00	35,22	48,31	53,52	59,21	57,71
4	0,56	20,03	25,71	39,32	46,6	56,5	57,81
5	0,41	31,61	34,83	46,51	46,7	48,5	51,11
6	0,45	22,01	28,11	38,5	46,71	54,11	53,62
7	0,34	25,71	31,4	41,1	50,6	53,5	49,31
8	0,74	18,7	26,8	37,8	50,6	65,02	72,31
9	0,75	27,32	34,61	47,81	55,91	67,90	75,20
10	0,48	18,31	22,81	32,81	43,41	49,60	51,21

U tablici 23. prikazane su svojstvene vrijednosti iz čega je vidljivo da prve tri komponente sadržavaju približno 99 % varijance u podatcima.

Tablica 23. Svojstvene vrijednosti (*program Statistica*)

Eigenvalues of covariance matrix, and related statistics (pcr) Active variables only				
	Eigenvalue	**% Total - variance**	**Cumulative - Eigenvalue**	**Cumulative - %**
1	209.7672	72.26273	209.7672	72.2627
2	73.8735	25.44870	283.6408	97.7114
3	4.6362	1.59712	288.2769	99.3086
4	0.9311	0.32077	289.2081	99.6293
5	0.7954	0.27401	290.0035	99.9033
6	0.2806	0.09667	290.2841	100.0000

Rezultat PCA prikazan je za prve tri komponente u tablici 24.

Tablica 24. Rezultat PCA (program *Statistica*)

Eigenvectors of covariance matrix						
	Factor 1	**Factor 2**	**Factor 3**	**Factor 4***	**Factor 5***	**Factor 6***
A1	-0.124184	-0.591943	-0.254197	-0.051539	0.339824	0.671885
A2	-0.017862	-0.512458	0.047373	0.190501	0.495660	-0.672944
A3	0.065636	-0.570855	-0.102121	0.136705	-0.791605	-0.118577
A4	0.243801	-0.239530	0.573483	-0.744455	-0.009393	-0.001356
A5	0.510160	-0.043144	0.545610	0.599932	0.065336	0.275678
A6	0.812558	0.043337	-0.544185	-0.168028	0.088573	-0.075206

**Faktori 4, 5 i 6 ne sudjeluju u daljnjem računu.*

Prve tri glavne komponente sadržavaju 99,30 % varijance u podatcima, stoga se nova regresijska analiza, umjesto s pomoću originalnih podataka, računa s pomoću prvih triju komponenti.

Vrijednosti skora (Z1, Z2 i Z3) izračunane su s pomoću podataka u tablicama 23. i 25.

Primjer je računa za prvu vrstu (Z1):

Z1=-0,124 x (A1) – 0,0178 x (A2) + 0,066 x (A3) + 0,0244 x (A4) + 0,510 x (A5) + 0,813 x (A6)

= -0,124 x (18,71) – 0,0178 x (26,80) + 0,066 x (42,21) + 0,0244 x (56,61) + 0,510 (70,20) + 0,813 x (83,30) = 117,38

a potom se na isti način izračunavaju vrijednosti za Z2 i Z3.

Vrsta	Z1	Z2	Z3
1 . . 10	117,38	-61,66	-17,78

Nova regresijska jednadžba nakon provedene PCA metode s trima glavnim komponentama glasi:

c = 0,067 + 0,011691 Z1 + 0,004070 Z2 – 0,016673 Z3.

Tablica 25. Rezultat regresije nakon provedene PCA (program *Statistica*).

Regression Summary for Dependent Variable: c (pcr) R= .99809023 R2= .99618410 Adjusted R2= .99427615 F(3,6)=522.12 p						
	Beta	**Std.Err. - of Beta**	**B**	**Std.Err. - of B**	**t(6)**	**p-level**
Intercept			0.067347	0.056812	1.18545	0.280656
z1	0.957878	0.025220	0.011691	0.000308	37.98088	0.000000
z2	0.199222	0.025232	0.004070	0.000515	7.89558	0.000219
z3	-0.204178	0.025233	-0.016673	0.002060	-8.09174	0.000191

Nakon provođenja PCA metode i ponovne regresije uz upotrebu izračunanih vrijednosti skora, pokazalo se da su sva tri koeficijenta statistički značajna (tablica 25.).

Uvrste li se vrijednosti Z1, Z2 i Z3 u novu regresijsku jednadžbu, dobiva se vrijednost koncentracije vrste.

Pitanja:

a) Zbog čega se u PCR-u regresija računa s glavnim komponentama, a ne izravno na originalnim varijablama?
b) Koja je razlika između PCA-a i PCR-a?
c) Što se događa s pojavom multikolinearnosti u PCR-u?
d) Koja komponenta najviše doprinosi predikciji zavisne varijable?
e) Koji se pokazatelji regresije poboljšavaju nakon PCR-a?

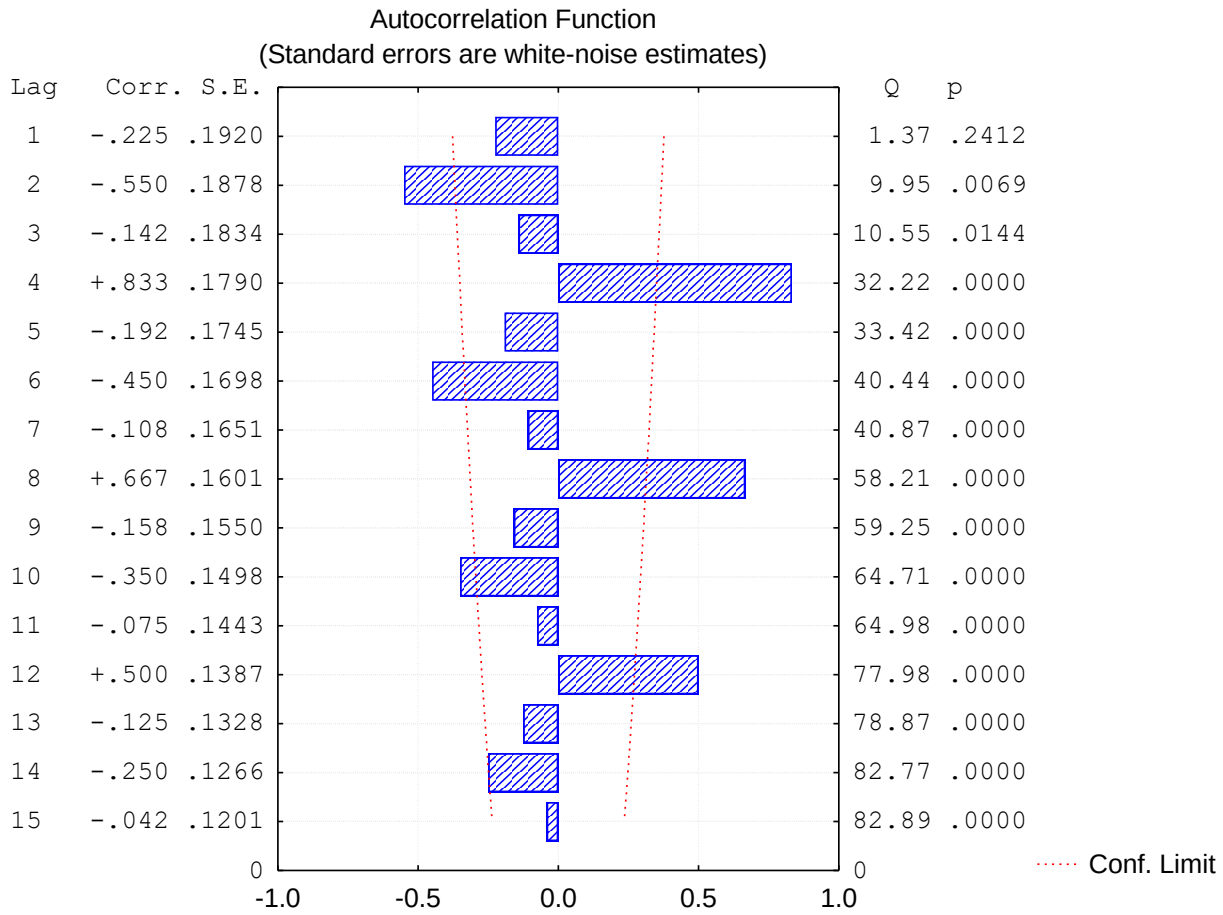
6. ANALIZE VREMENSKIH SERIJA

6.1. Autokorelacija

Autokorelacija opisuje stupanj u kojem je vrijednost neke varijable povezana s vlastitim prethodnim ili susjednim vrijednostima, odnosno njezinu korelaciju sa samom sobom kroz vrijeme ili prostor. U obradi velikih skupova podataka ta je metoda značajna za otkrivanje latentnih uzoraka i trendova koji nisu izravno uočljivi iz primarnih mjerenja. Primjerice, analiza vremenskih serija koncentracije metabolita u biološkim uzorcima može pokazati povezane cikluse ili oscilacije unutar sustava, dok u kemijskoj kinetici koncentracija produkata u određenom trenutku često izravno ovisi o vrijednostima iz prethodnih faza reakcije. Temeljni je koncept u toj analizi vremenski pomak (engl. *lag*) koji predstavlja broj razdoblja za koji se serija pomiče unatrag kako bi se ispitalo u kojoj mjeri trenutačno stanje sustava ovisi o njegovim prethodnim stanjima. Što je autokorelacija pri određenom pomaku veća, to je „memorija“ sustava snažnija. Posebno zanimljiva primjena autokorelacije nalazi se u praćenju onečišćenja zraka. Koncentracije tvari, poput čestica $PM_{2.5}$, NO_2, O_3 ili SO_2, u urbanim područjima često pokazuju snažnu vremensku autokorelaciju jer vrijednosti izmjerene u određenom satu ili danu mogu ovisiti o akumulaciji i zadržavanju tvari iz prethodnog razdoblja mjerenja. Detaljna analiza tih obrazaca omogućuje kemičarima i analitičarima dublje razumijevanje dinamike onečišćenja, preciznije predviđanje budućih trendova te razvoj optimiziranih strategija za smanjenje emisija i zaštitu okoliša.

Primjer: Analiza cikličkih oscilacija koncentracija PM2.5 čestica

Iz autokorelacijskog dijagrama prikazanog na slici 27., koji prikazuje autokorelacijsku funkciju koncentracija PM2.5 čestica, vidljivi su ciklusi od četiri dana. Ti ciklusi upućuju na periodične oscilacije u koncentracijama PM2.5 čestica koje mogu biti povezane s meteorološkim uvjetima (brzina vjetra, smjer vjetra, temperatura) ili antropogenim utjecajima (ložišta, promet i industrijske emisije).



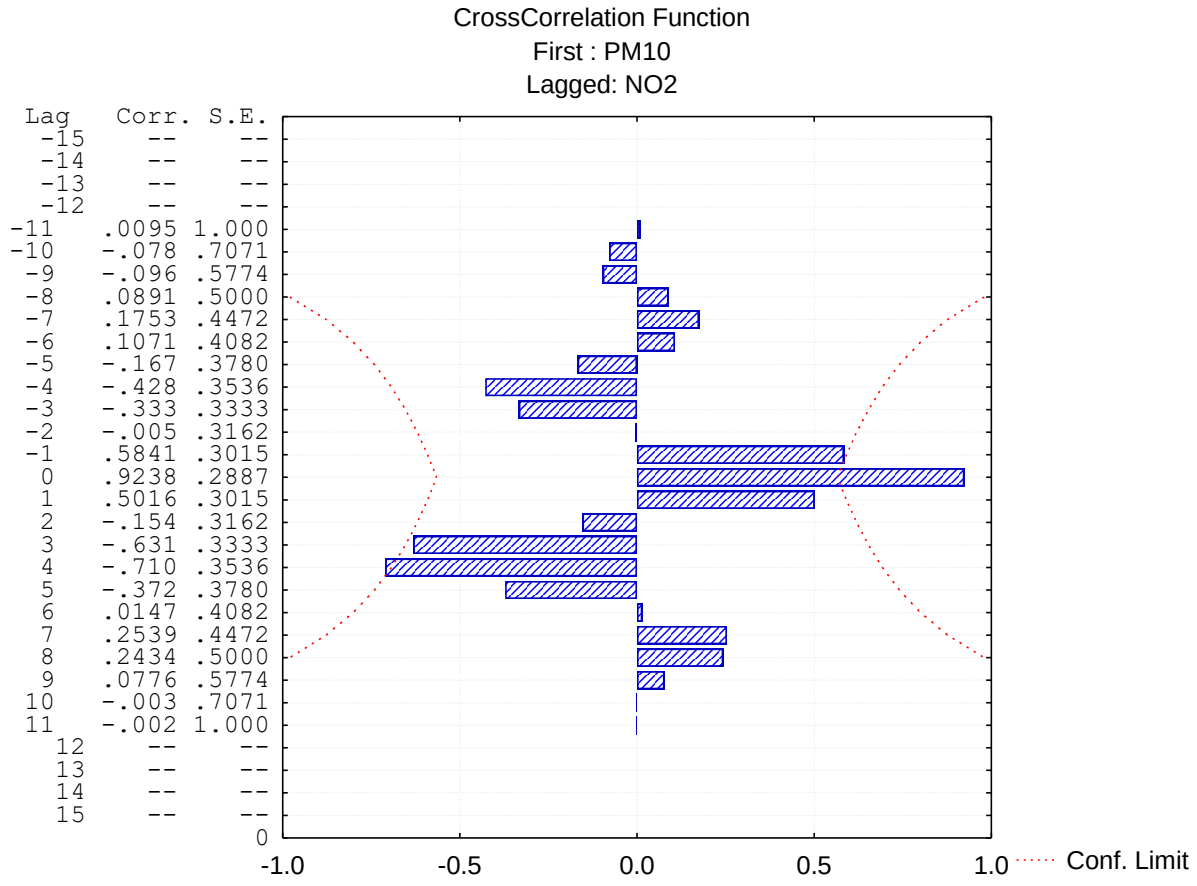
Slika 27. Autokorelacijski dijagram podataka koncentracija PM2,5 mjerenih 24 dana (program *Statistica*)

6.2. Kroskorelacija

Kroskorelacija je metoda kojom se ispituje povezanost između dviju vremenskih serija kada jedna može „zaostajati" ili „prednjačiti" u odnosu na drugu. Za razliku od standardne korelacije, koja pokazuje povezanost samo u istom trenutku, kroskorelacija omogućuje uočavanje dinamičkih utjecaja jedne serije na drugu kroz određeno razdoblje. Ta analiza pomaže prepoznati vremenska kašnjenja pri kojima je veza između dviju serija najjača. Primjenjuje se u mnogim područjima, poput meteorologije, kemije, biologije, medicine i drugih znanosti, gdje je cilj razumjeti odnose među pojavama koje se mijenjaju kroz vrijeme.

Primjer: Analiza vremenskog kašnjenja između PM10 i NO_2

Istražena je dinamička povezanost između dviju vremenskih serija: koncentracije lebdećih čestica (PM10) i koncentracije dušikova (II) oksida u urbanom zraku tijekom zimskog razdoblja. S obzirom na to da emisije tih onečišćivača često potječu iz istih izvora, poput ložišta i prometa, ali se različitom brzinom šire i zadržavaju u atmosferi, cilj je primijeniti metodu kroskorelacije kako bi se utvrdilo postoji li karakteristično vremensko kašnjenje u njihovim maksimalnim koncentracijama.



Slika 28. Kroskorelacija između PM10 i NO_2 (program *Statistica*)

Slika 28. prikazuje kroskorelacijsku funkciju PM10 i NO_2 u razdoblju od 24 sata pri čemu su koncentracije obaju onečišćivača mjerene svaki sat. Pri vrijednosti pomaka (engl. lag) = 0, kroskorelacijska funkcija pokazuje koliko su dvije promjene istovremeno povezane bez kašnjenja.

Najviša korelacija (0,9238) zabilježena je pri pomaku = 0 sati što upućuje na istovremene promjene obaju onečišćivača, ali i na zajednički izvor emisije (npr. promet, industrija). Pozivni pomaci upućuju na to da promjene u koncentracijama NO_2 mogu prethoditi promjenama u koncentracijama PM10, a negativni pomaci upućuju na obrnutu dinamiku.

6.3. Fourierove transformacije

U dugačkim nizovima mjerenja, poput meteoroloških parametara ili koncentracija onečišćivača u atmosferi, često su prisutni periodički obrasci koji nisu neposredno uočljivi izravnim promatranjem sirovih podataka. Ti skriveni ciklusi, koji se najčešće manifestiraju kao dnevni, tjedni ili sezonski ciklusi, predstavljaju ključne komponente dinamike ekoloških i fizikalno-kemijskih sustava.

Fourierova transformacija predstavlja temeljni matematički alat za otkrivanje i određivanje takvih periodičkih pojava. Temeljna pretpostavka te metode jest da se svaki složeni signal može pretvoriti u sumu jednostavnih periodičkih funkcija, odnosno sinusnih i kosinusnih valova različitih amplituda i frekvencija. Primjenom te pretvorbe podatci se iz vremenske domene prevode u frekvencijsku domenu.

Glavna prednost takva pristupa jest mogućnost preciznog određivanja dominantnih frekvencija unutar signala te procjena njihove relativne snage. Na taj način Fourierova transformacija omogućuje detektiranje i kvantificiranje cikličkih procesa koji ostaju skriveni u standardnom vremenskom prikazu pružajući dublji uvid u periodičku prirodu analiziranih pojava. Fourierova transformacija može se izračunati prema izrazu (19):

$$X_k = \sum_{n=0}^{N-1} x_n \cdot e^{-2\pi i k n/N} \tag{19}$$

gdje su: $k = 0,1,2\ldots., N-1$, X_k je amplituda frekvencije k, X_n *je* vrijednost signala u trenutku n, a *N* = broj mjerenja.

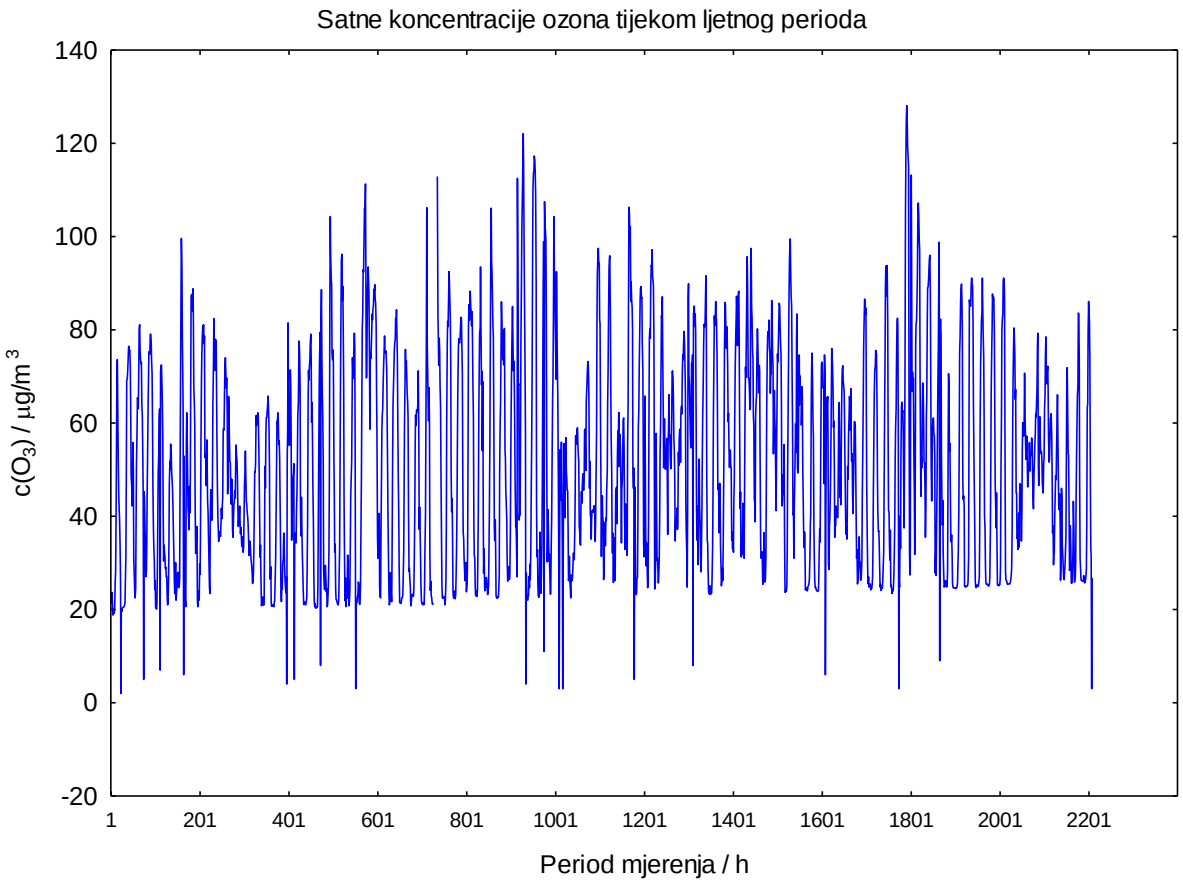
U vremenskim serijama prate se vrijednosti koncentracija tvari u zraku u nizu vremenskih točaka, npr. ozona (O_3), dušikovih oksida ($NO + NO_2 = NO_x$), sumporova (IV) oksida (SO_2) ili lebdećih čestica (PM2.5, PM10). Takvi podatci često skrivaju cikluse koji nisu odmah vidljivi u običnom grafikonu, primjerice u vremenskoj domeni u nizu podataka mjerenih kroz dulje razdoblje signal izgleda kaotično i ciklusi se ne vide, ali u frekvencijskoj domeni Fourierov spektar jasno pokazuje vrhove. Fourierova transformacija razbija signal na osnovne oscilacije i otkriva *nevidljive* cikluse. Koncentracije ozona i onečišćivača često pokazuju satne, dnevne, tjedne i sezonske cikluse zbog utjecaja sunčeve radijacije, prometa, industrije i meteoroloških uvjeta na vrijednosti njihovih koncentracija. Analiza vremenskih serija omogućava prepoznavanje periodičnih ciklusa u onečišćenju zraka, razlikovanje cikličkih promjena od slučajnih ili iznenadnih onečišćenja kao i predviđanje razine onečišćenja u budućnosti.

Mjere li se koncentracije ozona svaki sat ili svaki dan, grafikon izgleda kaotično, ali Fourierova analiza može otkriti različite skrivene cikluse uključujući i tjedni ciklus poznat pod nazivom vikend-efekt (engl. *weekend effect*). Rezultat je spektralni grafikon na čijoj se *x* osi nalazi frekvencija (npr. broj ciklusa po danu, mjesecu ili godini), a na *y* osi jakost ili amplituda ciklusa. Najizraženiji vrh u spektru označava dominantan ciklus u podatcima o onečišćenju.

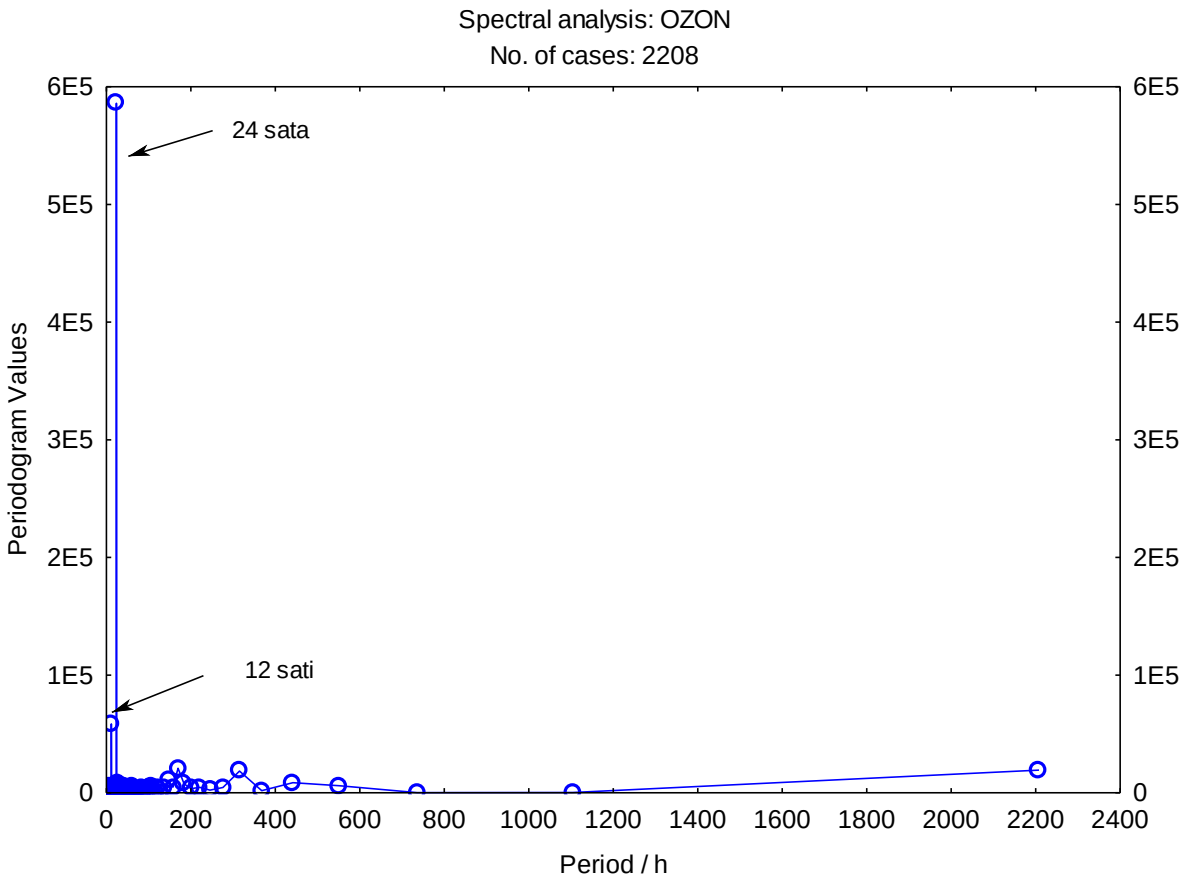
6.4. **Zadatci**

6.4.1. Na temelju prikazanih rezultata (slike 29., 30. i 31.) analizirajte dinamiku satnih koncentracija troposferskog ozona mjerenih tijekom ljetnog razdoblja u urbanom području. Na slici 29. prikazan je grafikon satnih koncentracija zona mjerenih tijekom ljetnog razdoblja na području grada. Skriveni ciklusi nisu uočljivi na grafikonu sve dok se podatci ne podvrgnu metodi Fourierovih transformacija.

Na slici 30. prikazan je periodogram na kojem je jasno vidljiv dominantni ciklus na 24 sata kao i nešto slabije izražen ciklus na 12 sati.

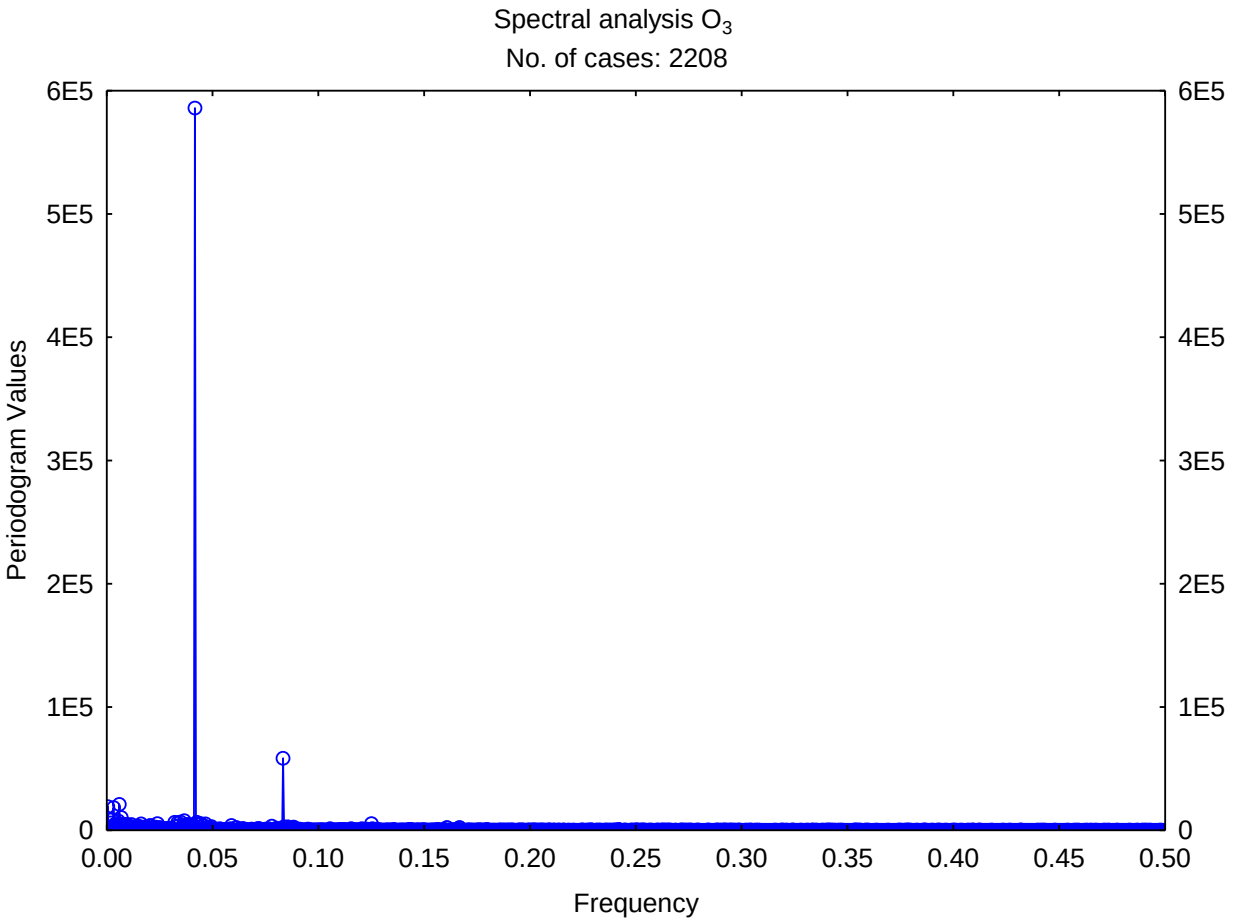


Slika 29. Periodogram satnih koncentracija O_3 (µg/m^3) (program *Statistica*)



Slika 30. Periodogram (program *Statistica*)

Ciklusi postaju jasnije uočljivi prikaže li se rezultat na drugi način (slika 31.), (program *Statistica*).



Slika 31. Periodogram za O_3 (*program Statistica*)

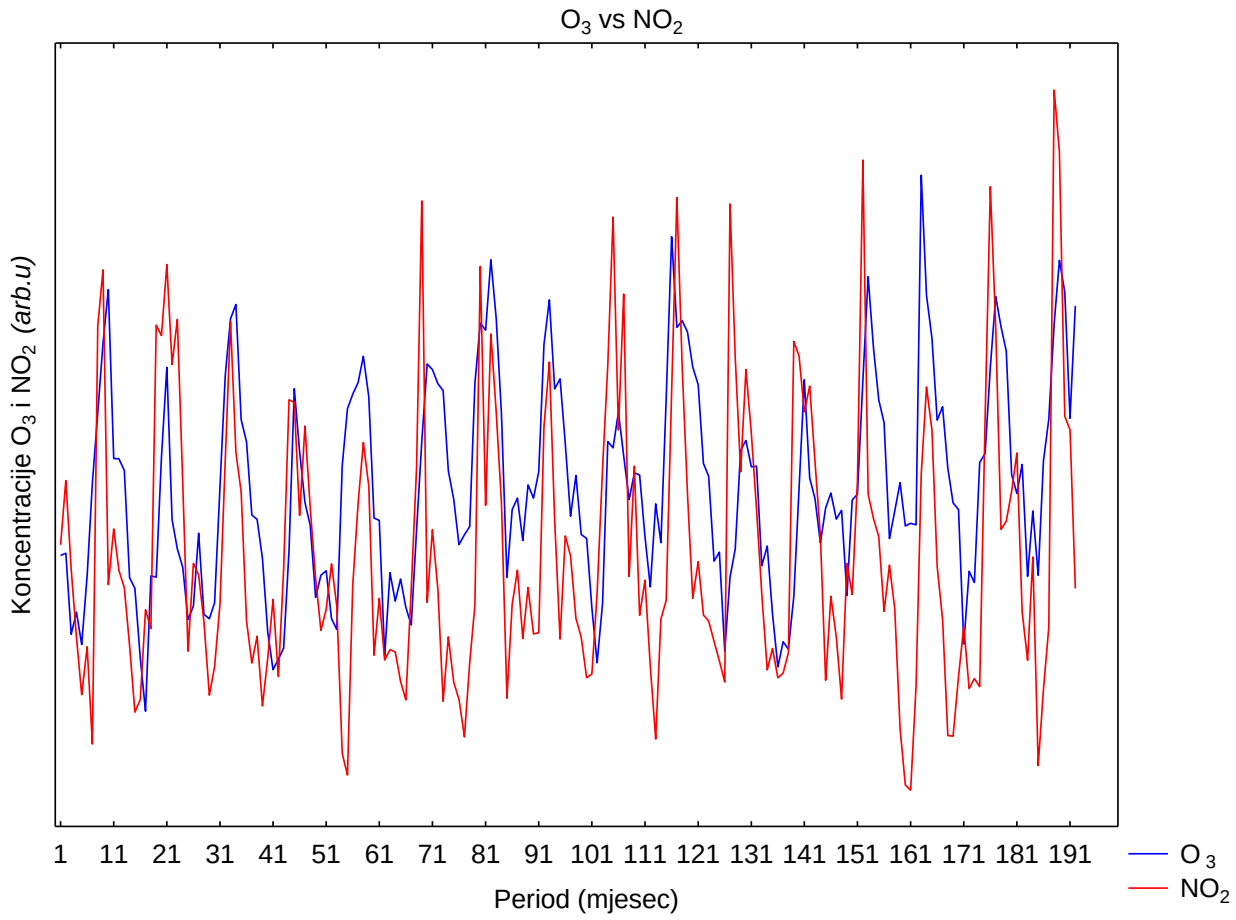
Pitanja:

a) Koji vremenski ciklus prikazuje najintenzivniji pik na slici 31.?

b) Koje vremenske cikluse prikazuju ostali pikovi manjeg intenziteta?

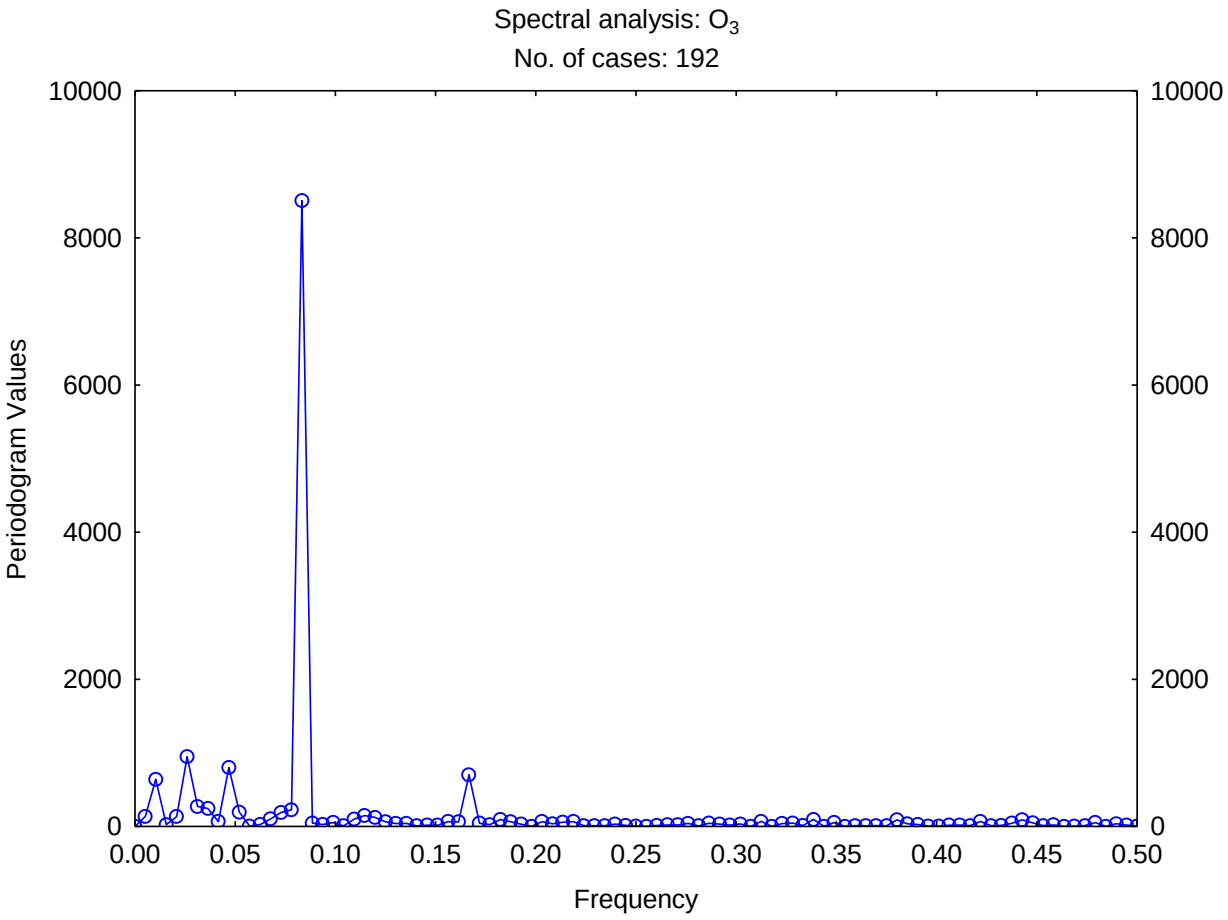
c) Koji način prikaza rezultata daje izraženije pikove?

6.4.2. Na podatke koncentracija ozona i NO_2 (slika 32.) primijenjena je Fourierova analiza. Objasnite dobivene rezultate s obzirom na dinamiku emisija i kemijsku transformaciju prekursora u atmosferi.

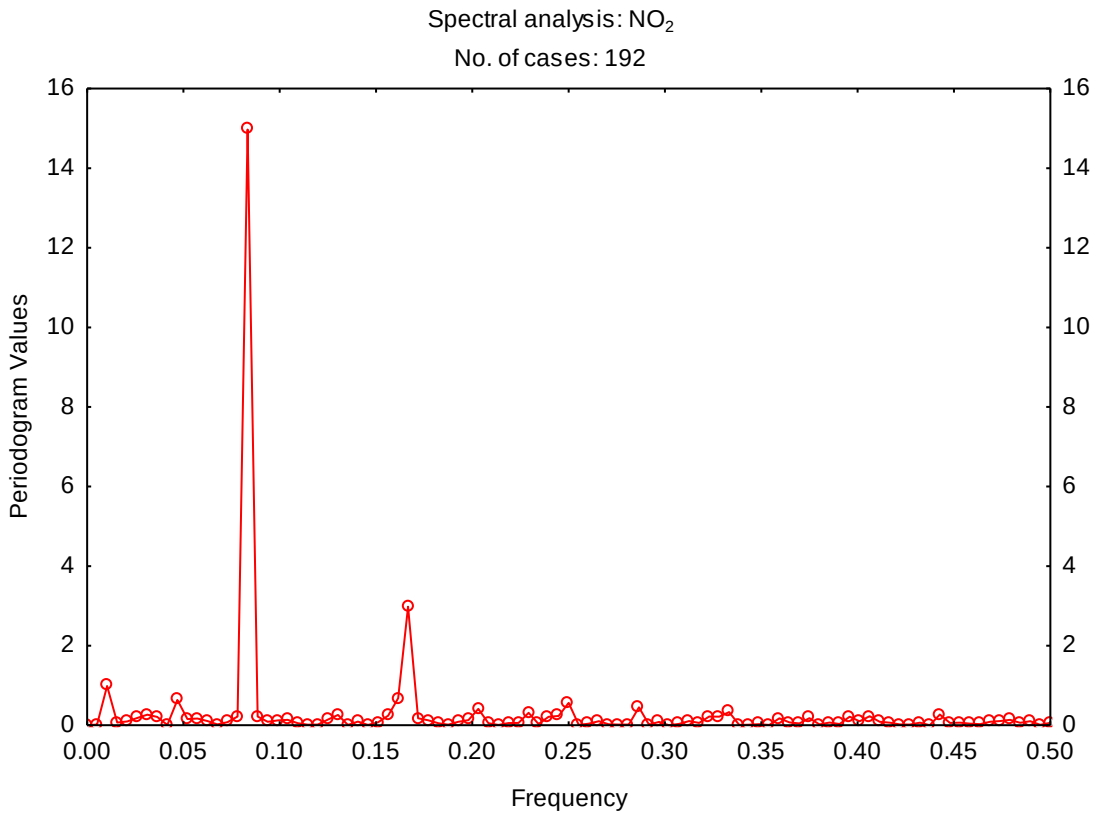


Slika 32. Mjesečni prosjeci koncentracija O_3 i NO_2 na području Brazila (program *Statistica*)

Rezultati Fourierove analize za O_3 i NO_2 prikazani su na slikama 33. i 34.



Slika 33. Periodogram za O_3 (program *Statistica*)



Slika 34. Periodogram za NO_2 (program *Statistica*)

Pitanja:

a) Na temelju usporedbe periodograma prikazanih na slikama 33. i 34. interpretirajte ključne razlike u spektralnim obilježjima analiziranih signala.

b) Koja je primarna svrha Fourierove transformacije u analizi vremenskih serija?

c) Što prikazuje periodogram?

d) Koji kemijski, biološki ili meteorološki ciklusi odgovaraju frekvencijama u periodogramu?

e) Obrazložite važnost prepoznavanja cikličkih obrazaca u koncentracijama atmosferskih onečišćivača za razumijevanje dinamike ekoloških sustava.

f) Kako biste objasnili podrijetlo dvaju onečišćivača ako su dva periodograma slična, odnosno ako imaju zajedničke cikluse?

g) Kako se periodogram koristi u predviđanjima kvalitete zraka?

7. DISKRIMINACIJSKA ANALIZA

Diskriminacijska je analiza multivarijantna metoda koja se koristi za razvrstavanje u unaprijed definirane grupe na temelju više nezavisnih varijabli. Cilj je metode pronaći linearnu kombinaciju varijabli koje u najvećoj mjeri razlikuju promatrane grupe i omogućavaju klasifikaciju daljnjih opažanja. Najčešće se primjenjuje linearna diskriminacijska analiza (LDA) koja pretpostavlja normalnu raspodjelu podataka i jednaku kovarijancu skupina. Ako te dvije pretpostavke nisu zadovoljene, koristi se kvadratna diskriminacijska analiza (QDA) koja dopušta različite kovarijacijske strukture. Osnovni cilj diskriminacijske analize jest pronaći manji broj latentnih varijabli koje se dobiju linearnom kombinacijom izvornih, a nazivaju se diskriminacijskim funkcijama.

Postupak diskriminacijske analize započinje odabirom relevantnih varijabli i provjerom osnovnih pretpostavki metode, nakon čega se računaju diskriminacijske funkcije koje maksimalno odvajaju promatrane skupine. Značajnost dobivenih funkcija procjenjuje se statističkim testovima (Wilksova lambda). Na temelju dobivenih funkcija moguće je klasificirati promatrane jedinice i procijeniti točnost svake klasifikacije.

Diskriminacijska analiza ima široku primjenu u različitim područjima: kemiji, biologiji, medicini, društvenim znanostima itd.

Nakon što se postupak diskriminacijske analize provede na uzorcima za koji je pripadnost skupinama poznata, dobiva se jedna *diskriminacijska funkcija* ili više njih. Te funkcije predstavljaju matematičke modele kojima su opisane razlike među skupinama na temelju odabranih varijabli. Svaka skupina ima svoj centroid, tj. prosječnu vrijednost diskriminacijske funkcije. Kada se pojavi nepoznati uzorak, u njega se uvrste izmjerene vrijednosti istih varijabli i izračunava se vrijednost diskriminacijske funkcije, a dobivena se vrijednost uspoređuje s centroidima poznatih skupina. Uzorak se svrstava u skupinu čijem je centroidu najbliži, odnosno u skupinu za koju ima najveću vjerojatnost pripadnosti.

Na taj način diskriminacijska analiza omogućava objektivno i statistički utemeljeno razvrstavanje novih opažanja.

7.1. **Zadatci**

7.1.1. Potrebno je razvrstati džemove dobivene iz triju vrsta jabuka (zelenih, crvenih i žutih) na temelju koncentracija: glukoze, fruktoze, saharoze i florizina. Potrebno je izračunati u koju će se skupinu razvrstati **novi** džem koji sadržava: 10, 9, 23, 50 i 3,851 mg/L saharoze, glukoze, fruktoze i florizina.

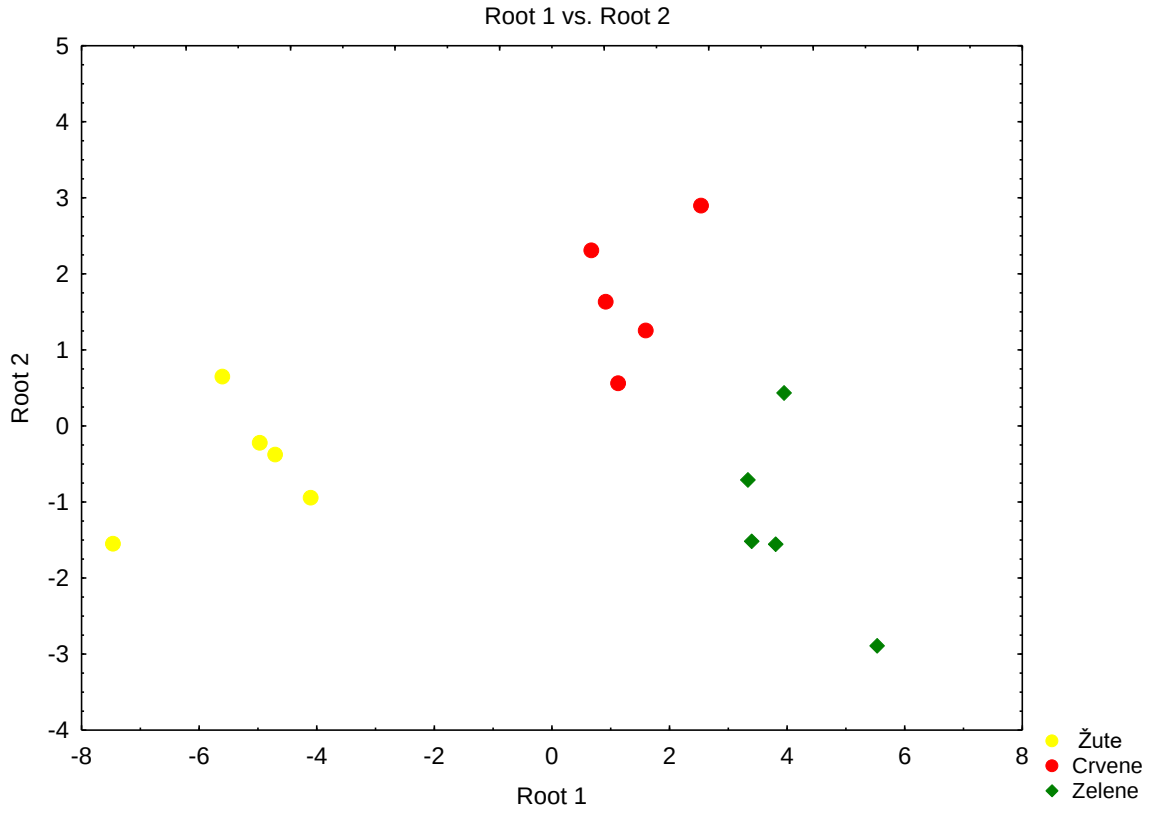
Tablica 26. Koncentracije saharoze, glukoze, fruktoze i florizina (mg/L)

Boja jabuke	Saharoza	Glukoza	Fruktoza	Florizin
žute	34	5	47	2,9
žute	29	16	40	7,3
žute	6	26	49	3,8
žute	10	22	47	3,5
žute	14	21	51	6,3
crvene	10	20	49	3,3
crvene	8	19	49	3,5
crvene	8	17	55	5,3
crvene	7	21	59	3,3
crvene	15	20	68	4,9
zelene	14	19	74	5,6
zelene	9	15	57	5,4
zelene	34	5	47	2,9
zelene	29	16	40	7,2
zelene	6	26	49	3,8

Na temelju podataka tablice 26., u programu *Statistica* provedena je diskriminacijska analiza. Diskriminacijska funkcija prikazana je u tablici 27., a dijagram kanoničkih skorova prikazan je na slici 35. Kanonički skorovi pružaju uvid u to koje kombinacije varijabli najbolje objašnjavaju varijacije među grupama i mogu se koristiti za vizualizaciju razlika među grupama podataka.

Tablica 27. Diskriminacijska funkcija (program *Statistica*)

Variable	Classification Functions		
	A p=.33333	B p=.33333	C p=.33333
Saharoza	0.3876	-1.6633	-2.500
Fruktoza	1.4603	2.5275	3.478
Glukoza	0.4213	1.2056	0.541
Florizin	2.1899	3.5881	5.477
Constant	-45.2936	-75.3430	-115.113



Slika 35. Dijagram kanoničkih skorova (program *Statistica*)

Za novi džem linearne diskriminacijske vrijednosti skorova za svaku skupinu iznose:

žute: -45,2936 + 0,3876 · 10,9 + 1.46 · 50 + 0,42 · 23 + 2,19 · 3,85 = 49,89

crvene: -75,3430 – 1,6633…… dovršite račun

zelene: -115,113 – 2,500……. dovršite račun

Na temelju zadanih linearnih diskriminacijskih funkcija, izračunajte vrijednosti skorova za preostale dvije skupine (crvene i zelene). Usporedbom dobivenih vrijednosti odredite skupinu s najvišim skorom te, u skladu s pravilom klasifikacije, odredite kojoj sorti jabuka pripada novi uzorak džema koji sadržava: 10,9, 23, 50 i 3,85 mg/L saharoze, glukoze, fruktoze i florizina.

7.1.2. U tablici 28. prikazan je sastav minerala prema udjelu Ca, Mg, Si i K. Nakon provedene diskriminacijske analize, potrebno je utvrditi kojoj skupini pripada nepoznati uzorak.

Tablica 28. Klasifikacija minerala prema sastavu Ca, Mg, Si i K

Uzorak br.	Mineral	Ca (%)	Mg (%)	Si (%)	K (%)
1	kvarc	40,01	0,51	0,10	0,20
2	kvarc	41,03	0,42	0,20	0,11
3	dolomit	30,00	17,50	0.30	0,20
4	dolomit	28,01	17,01	0,20	0,30
5	magnetit	3,01	0,10	69,01	0,20
6	magnetit	3,02	0,21	68,01	0,20
Nepoznati uzorak	**?**	31,02	18,03	0,25	0,25

Pitanja:

1. Što je to linearna diskriminacijska funkcija?
2. Na koji se način interpretiraju koeficijenti linearne diskriminacijske analize?
3. Koje varijable najmanje doprinose razlikovanju skupina?
4. Koje biste dodatne analize dodali za bolje razlikovanje u obama navedenim primjerima (džem i minerali)?
5. Na koji način normalizacija varijabli utječe na linearnu diskriminacijsku analizu?

8. LITERATURA

1. I. Šošić, *Primijenjena statistika*, Školska knjiga, Zagreb, 2004.

2. E. Kovač Andrić, V. Gvozdić, B. Matasović, N. Sakač, A. de Souza, *Analysis of Stratospheric Ozone and Nitrogen Dioxide over Mid Brazil for a Period from 2005 to 2020*, Atmosphere**,** 16(10), 1159, 2025.

3. E. Kovač Andrić, J. Brana, V. Gvozdić, *Impact of meteorological factors on ozone concentrations modelled by time series analysis and multivariate statistical methods*, Ecological Informatics, 4(2), 117–122, 2009.

4. D. Bjedov, A. Mikuška, V. Gvozdić, P. Glavaš, D. Gradečak, M. Sudarić Bogojević, *White stork pellets: Non-invasive solution to monitor anthropogenic particle pollution*, Toxics, 12(4), 236, MDPI, Basel, 2024.

5. V. Gvozdić, E. Kovač Andrić, J. Brana, *Influence of Meteorological Factors NO_2, SO_2, CO and PM_{10} on the Concentration of O_3 in the Urban Atmosphere of Eastern Croatia*, Environmental Modeling & Assessment, 16(5), 491–501, 2011.

6 L. Massart, B. G. M. Vandeginste, S. Deming, Y. Michotte, L. Kaufman, *Handbook of Chemometrics and Qualimetrics, Part A*, Elsevier, Amsterdam, 1997.

7. L. Massart, B. G. M. Vandeginste, S. Deming, Y. Michotte, L. Kaufman, *Handbook of Chemometrics and Qualimetrics, Part B*, Elsevier, Amsterdam, 1998.

8 A. C. Rencher, W. F. Christensen, *Methods of Multivariate Analysis*, 3. izd., Wiley, 2012.

9. I. T. Jolliffe, *Principal Component Analysis*, 2. izd., Springer, 2002.

10. T. M. Powell, J. H. Steele (ur.), *Ecological Time Series*, Springer, New York, 1995.

11. J. N. Miller, J. C. Miller, *Statistics and Chemometrics for Analytical Chemistry*, 6th ed., Pearson Education Limited, Harlow, 2010.

12. P. Atkins, J. de Paula, Atkins' *Physical Chemistry*, 6th ed., Oxford University Press, Oxford, 2009.

13. M.S. Silberberg, *Chemistry:The Molecular Nature of Matter and Change*, 7th ed., Mc Graw-Hill Education, New York, 2015.

Kazalo pojmova

Adjusted R^2 – prilagođeni koeficijent determinacije

Autocorrelation function – autokorelacijska funkcija

Cluster Analysis – klasterska analiza

Confidence interval – interval pouzdanosti

Confidence limits – granice pouzdanosti

Cross-correlation function – kroskorelacijska funkcija

Eigenvalues of correlation matrix – svojstvene vrijednosti korelacijske matrice

Eigenvalue number – redni broj vlastite vrijednosti

Factor loadings – faktorska opterećenja glavnih komponenti

Factor scores – faktorski skorovi glavnih komponenti

Fourier Transform Infrared Spectroscopy (FT–IR) – infracrvena spektroskopija s Fourierovom transformacijom

Inductively Coupled Plasma Mass Spectrometry (ICP-MS) – spektrometrija masa uz induktivno spregnutu plazmu

Intercept – odsječak

Lag – pomak

QSAR, Quantitative Structure-Activity Relationship – kvantitativni odnos strukture i aktivnosti

QSPR, Quantitative Structure-Property Relationship – kvantitativni odnos strukture i svojstava

Standard Error of Estimate (SE) – standardna pogreška procjene

Predicted Residual Sum of Squares (PRESS) – prediktivna rezidualna suma kvadrata

Principal Component Analysis (PCA) – analiza glavnih komponenti

Principal Components (PC_s) – glavne komponente

Principal Component Regression (PCR) – regresija s glavnim komponentama

PCA scores – skorovi glavnih komponenti

Root Mean Squared Error (RMSE) – korijen srednje kvadratne pogreške

Scree plot – dijagram loma